Article:
Shao, Z., Han, J., Debattista, K. et al. (1 more author) (2024) DCMSTRD: End-to-end dense captioning via multi-scale transformer decoding. IEEE Transactions on Multimedia, 26. pp. 7581-7593. ISSN 1520-9210

https://doi.org/10.1109/TMM.2024.3369863

DCMSTRD: End-to-end Dense Captioning via Multi-Scale Transformer Decoding

Zhuang Shao, Jungong Han, Kurt Debattista, Yanwei Pang

***Abstract*—Dense captioning creates diverse Region of Interests (RoI) descriptions for complex visual scenes. While promising results have been obtained, several issues persist. In particular: 1) it is hard to find the optimal parameters for artificially designed modules (e.g., non-maximum suppression (NMS)) causing redundancies and fewer interactions to benefit the two subtasks of RoI detection and RoI captioning; 2) the absence of a multi-scale decoder in current methods hinders the acquisition of scale-invariant features, thus leading to poor performance. To tackle these limitations, we bypass the artificially designed modules and present an end-to-end dense captioning framework via multi-scale transformer decoding (DCMSTRD). DCMSTRD solves dense captioning by set matching and prediction instead. To further enhance the discriminative quality of the multi-scale representations during caption generation, we introduce a multi-scale module, termed multi-scale language decoder (MSLD). Our proposed method tested on standard datasets achieves a mean Average Precision (mAP) of 16.7% on the challenging VG-COCO dataset, demonstrating its effectiveness against the current methods.**

***Index Terms*—Dense Captioning, Artificially Designed Modules, End-to-end Dense Captioning framework via Multi-Scale Transformer Decoding (DCMSTRD), Multi-Scale Language Decoder (MSLD)**

I. Introduction

Dense captioning is an extension of image captioning [1]. Instead of producing a single caption for the entire image, dense captioning aims to detect all the Region of Interests (RoIs) in the input and describe them via natural language. Thanks to the salient-part descriptors that provide rich and dense semantic visual information, dense captioning can benefit other tasks, including visual question answering [2], image segmentation [3], [4], and action recognition [5], [6].

Most current image captioning methods adhere to an encoder-decoder architecture, a paradigm prompted by the successful transfer of sequence-to-sequence training to achieve machine translation [7]. In this architecture, a Convolutional Neural Network (CNN) typically acts as an encoder to extract image features before they are decoded by a trainable Recurrent Neural Network (RNN). Yet, the simplicity of these encoder-decoder frameworks leads to descriptions that do not focus well on salient regions and object information in an image. To remedy such issues, subsequent work focused on designing different weight modules for each feature map. Specifically, [8], [9] implemented a Bottom-Up and Top-Down Attention algorithm, prioritizing different regions from the learned weights of feature maps. Likewise, [10] devised a dual-stream co-attention module to combine different kinds of visual features to produce the corresponding different words. Although these approaches have achieved relatively good performances, further works have explored diverse architectures and other unique evaluation metrics. This exploration has been propelled along two intersecting trajectories. Firstly, Transformer [11] frameworks were used to contribute to the generation of image captions. For example, [12] proposed an Attention on Attention model, which extended the self-attention in the Transformer to optimise the results of attention. [13] proposed a Transformer-based structure to align grid features to reduce semantic noise in attention. Additionally, [14] introduced RSTNet, an architecture combining spatial information and adaptive attention to bridge the gap between non-visual signals and textual content. Secondly, recent strides in image captioning have enhanced the diversity and distinctiveness of the generated captions. [15] proposed a framework with latent spaces of context-object split to create more diverse captions. In a similar vein, [16] presented a novel metric named *CIDErBtw* to supervise the training, elevating the distinctiveness of images sharing similar themes.

Generally, *dense captioning* emerges as a more intricate task, compared to image captioning, due to the demanding need for deep comprehension of visual scenes and the generation of coherent natural language sequences for each region of interest. [20] set the trajectory for dense captioning and introduced an architecture named Fully Convolutional Localization Network (FCLN), which constitutes a bilinear interpolation location module to detect the RoIs and an LSTM decoder to produce the descriptions. Following this pioneering work, many alternatives were proposed, which can be roughly classified into two classes: context methods and non-context methods. Initially, architectures were centred around the adoption of Faster R-CNN [21] for RoIs localization and Long Short-Term Memory (LSTM) module [22] for the descriptors. This category aligns with so-called non-context methods, which processed the RoIs independently but did not leverage contextual knowledge for improved performance. Addressing this challenge, [23] pioneered an approach that combined RoI features with image features. This integration can be seen as creating a global context by fusing these

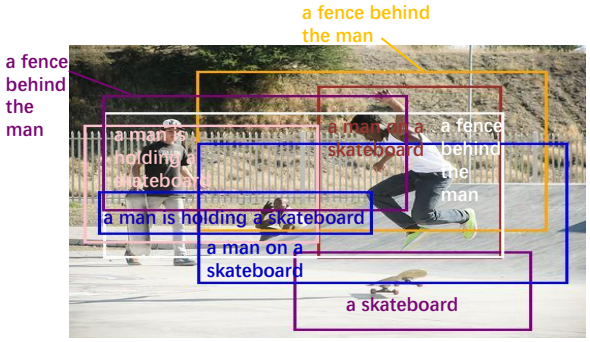
Manuscript received xxx, xxx; revised xxx, xxx and xxx, xxx; accepted xxx, xxx. (Corresponding author: Jungong Han). This research was supported by the funds of China Scholarship Council under Grant No. 201909120012.

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

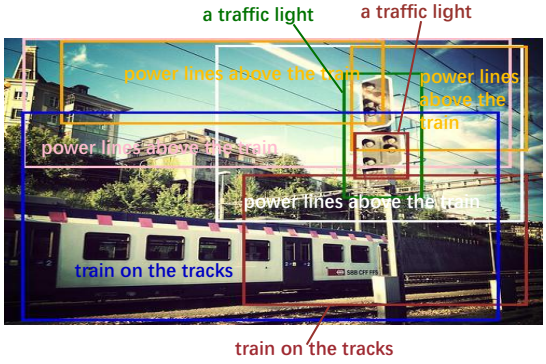
Jungong Han is with the Department of Computer Science, University of Sheffield, S1 4DP, UK (e-mail: jungonghan77@gmail.com).

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, and is also with Shanghai Artificial Intelligence Laboratory, China (e-mail: pyw@tju.edu.cn).

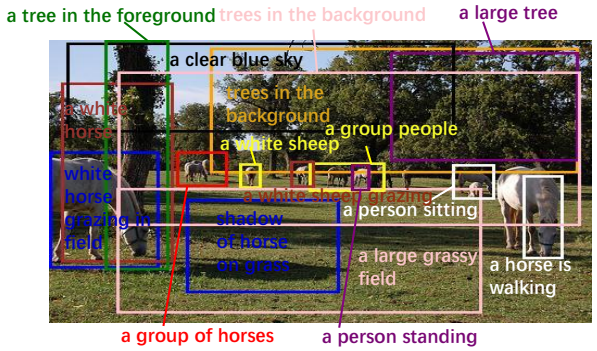


(a) The first example of visualisation results for dense captioning from [17].



(b) The second example of visualisation results for dense captioning from [18].

Fig. 1: Current methods always rely on artificially designed components especially non-maximum suppression (NMS), which always render redundancies and fewer interactions to benefit the two sub-tasks of RoI detection and RoI captioning. Two examples of dense captioning visualization from [18] and [17]. (The thresholds of pre-NMS and post-NMS are 0.7 and 0.5 respectively)



(a) The first example of visualisation results for dense captioning from [18].

(b) The second example of visualisation results for dense captioning from [19].

Fig. 2: Previous dense captioning methods always ignore the learning of multi-scale features and therefore fail to learn scale-invariant features for the same class of RoIs but with different scales. Two examples of dense captioning visualization from [18] and [19].

elements prior to captioning via an LSTM decoder. Nevertheless, this kind of context is too coarse, offering imprecise cues during training. To alleviate this phenomenon, several methods have been devised, delving into more refined forms of contextualization. For instance, [24] integrated the global, neighbours and target RoI features into a non-local similarity graph for caption generation. Alternatively, with the supportive data statistics, [19] extracted the object-level knowledge as context clues and transformed them into descriptions forming a close relationship between RoIs and detected objects via object detection in the whole architecture. Inspired by the Transformer architecture, Shao et al. [17] applied a Transformer model to encode objects and regions and devised a special loss (ROCSU) to exploit the region-object correlation score and improve the captioning performance. Furthermore, [18] presented a Textual Context Module (TCM) to capture the textual context and a Dynamic Vocabulary Frequency Histogram (DVFH) re-sampling framework to improve the word-learning efficiency during the training stage.

Despite the previous overall success of the aforementioned methods, the realization of the dense captioning system still remains imperfect. We believe that limitations still exist, with two, in particular, standing out as areas ripe for improvement. Firstly, existing methods [19] [24] [17] [18] predominantly adhere to a two-stage process, leveraging Faster R-CNN as their RoI detection frameworks. However, Faster R-CNN relies on an array of artificially designed modules, in particular, non-maximum suppression (NMS). These artificially designed modules usually include pre-defined parameters, but it is extremely difficult to find the optimal combination of these parameters for every image across the whole dataset. Hence, these modules tend to induce redundancies while hampering interactions to benefit the two sub-tasks of RoI detection and RoI captioning. We display this limitation via two examples chosen from the state-of-the-art methods [18] and [17] respectively in Fig. 1.

Fig. 1a demonstrates the challenge of choosing optimal parameters for the NMS threshold and other artificially designed components to effectively merge similar RoIs with a high Intersection of Union (IoU) [21] value for all the images across the dataset. This intricacy gives rise to repetitive captions. Specifically, ‘a fence behind the man’ appears four times in the results, while ‘a man is holding a skateboard’ and ‘a man on a skateboard’ are also repetitive (appears twice). This also

applies to Fig. 1b, in which 'power lines above the train' is repeated four times and 'a traffic light' and 'train on the tracks' repeat twice as a result of the poorly pre-defined threshold of the NMS components for this image, which fails to merge these repetitive results together.

Secondly, the decoders of previous dense captioning architectures are incapable of capturing multi-scale features, thus struggling a lot to learn scale-invariant features for objects of the same class but varying scales. As a result, the decoder is likely to produce poor captions for such RoIs or even ignore a part of multi-scale RoIs, leading to suboptimal performance. Examples of this shortcoming are shown in Fig. 2. In Fig. 2a, the decoder of [18] successfully captioned the two horses at the front (with a similar size, relatively large) whereas it fails to supply satisfactory captions for the differently sized horses in the background. It either recognizes the horses as sheep (the yellow RoI on the left with a caption of 'a white sheep' and the brown RoI in the middle with a caption of 'a white sheep grazing') or person or people (the white RoI with a description of 'a person sitting' on the right, the purple one in the middle 'a person standing' and the yellow RoI 'a group of people'). Furthermore, in Fig. 2b, even if the decoder of [19] roughly detected and described the RoIs (red and yellow) in the bottom right corner, it is unable to detect any of the persons in the background due to the missing of the multi-scale feature module and only ends up with the large purple RoI and an ambiguous sentence 'a large group', which is less accurate.

The main contributions of this paper are three-fold:

- We present an end-to-end dense captioning framework via multi-scale transformer decoding (DCMSTRD), which parallelizes the RoI selection and the captioning task. DCMSTRD solves the dense captioning task by set matching and prediction, instead of relying on artificially designed modules, thus alleviating the redundancy issue of conventional frameworks of prior works.
- A novel language module, named multi-scale language decoder (MSLD), is proposed to boost the learning of the multi-scale representations during the caption generation. The most significant feature of MSLD lies in its capacity to incorporate multi-scale features and supply more scale-invariant features with RoIs consisting of objects with the same class. This new module improves the accuracy of the caption generation.
- A thorough validation of our proposed DCMSTRD method on VGCOCO, VG V1.0 and VG V1.2 datasets demonstrates a substantial performance improvement over existing methods in terms of mean Average Precision (mAP).

The following part of this paper is organized as follows: To begin with, we review the prior works in Section II. Later on, in Section III, we present the proposed methodology and expound on the details of our presented DCMSTRD. Experimental results are displayed in Section IV with qualitative and quantitative discussions. Finally, we draw a conclusion and discuss our potential future works in Section V.

II. Related Work

A. Image Captioning

Image captioning [25]–[27] aims to generate descriptions based on the provided image. In the early stages of research, many solutions relied on retrieval-based methods. These approaches involved creating predefined templates within retrieval caption candidates [28] using simple visual feature encoders [8]. Nevertheless, a limitation of this approach was its inability to connect image descriptions with entirely new objects across the entire dataset [28]. To address this challenge, the field shifted towards deep learning methods as the technology and hardware advanced. The adoption of deep neural networks (DNN) became prevalent in the later stages of development. Initially, [29] introduced an embedding model of both image and text to build up a multi-modal sentence production model, utilizing a Convolutional Neural Network (CNN) as the encoder and a Long Short-Term Memory (LSTM) as the decoder, as well as [1] further proposed a reinforcement learning framework. In a later work, [30] designed a fine-grained region feature extractor using an R-CNN object detector [31], enabling the generation of region-level captions for the given image.

These encoder-decoder architectures assigned equal importance to all detected regions, neglecting the significance of certain regions that could provide crucial visual cues for captioning. To address this concern, diverse attention mechanisms were introduced due to their adaptable nature. [9] introduced a model incorporating semantic attention, whilst [32] introduced a bottom-up and top-down attentive module to represent the input image with a set of objects detected by a fixed object detector. In recent years, the adoption of Transformer regime [11] has greatly influenced Natural Language Processing (NLP) and various computer vision tasks including image captioning. [33] pioneered and proposed a Transformer pipeline for image captioning. This model extracted the whole image feature and uniformly sampled features by dividing the image into patches, which were input sequentially into the Transformer encoder [11] one by one. Several other Transformer-based solutions have also focused on enhancing captioning performance by exploring hidden properties. Among these efforts, [13] incorporated grid features to coordinate with Region of Interest (RoI) features, reducing semantic noise in the attention mechanism of the older Transformer architecture. Meanwhile, [14] presented RSTNet, a model that incorporated spatial information to flatten grid features and employed adaptive attention to connect textual and non-visual cues. On a different note, some prior research have sought to improve the variety and uniqueness of generated captions. Notably, [15] introduced separate latent spaces of context and objects to increase caption diversity. In another approach, [16] introduced a novel evaluation metric, namely CIDErBtw, which supervised the training process to generate distinctive words for images sharing common themes.

B. Dense Captioning

Despite the ability to supply a general overview of an entire image by image captioning, over time, it became evident that

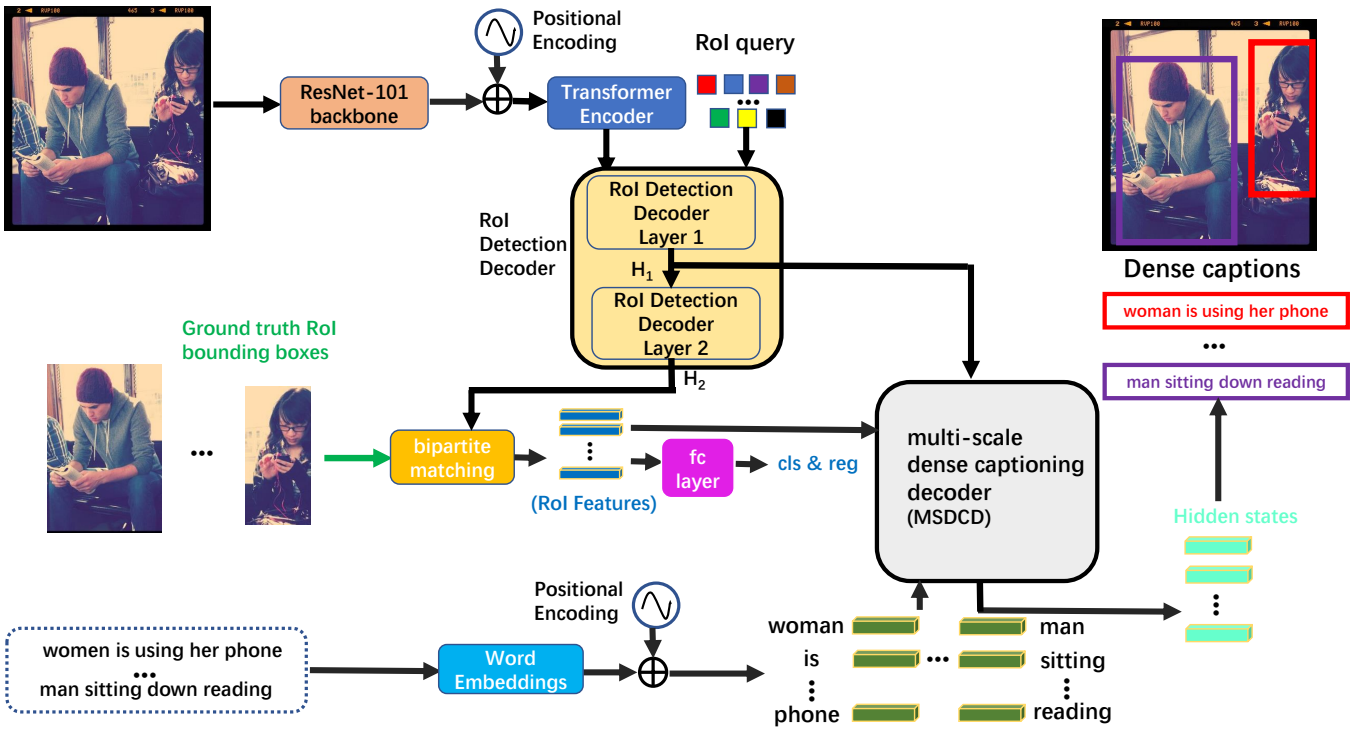


Fig. 3: The proposed DCMSTRD framework is made up of a Transformer Encoder, followed by an RoI detection detector, and a novel multi-scale language decoder (MSLD). Provided an image, the Transformer encoder takes the features extracted from the backbone and encodes them with the attention mechanism. The output and the RoI query are fed into the RoI detection decoder made up of two layers and decoded by the mutual attention mechanism inside. The outputs of these two layers are connected by the fc layer to attain the results of classification and bounding box regression before being matched with groundtruth by a bipartite matching method (Hungarian algorithm). Finally, the output of the two layers of the RoI detection decoder and the word embeddings are input into the novel multi-scale language decoder (MSLD) to capture multi-scale information and sentence information is decoded thus generating dense captions for each RoI.

this approach fell short of achieving more intricate and detailed descriptions. To remedy this, dense captioning [20] emerged as a novel task, demanding an intelligent vision system capable of both localizing and describing multiple significant regions within an image using natural language [18].

Due to the richer and fine-grained local descriptions, dense captioning can facilitate many industrial applications, such as blind navigation and human-robot interaction, and autonomous driving. To be specific, for blind navigation, after capturing an image using a smartphone camera and uploading it, the data is transferred to a distant server through a cloud system. In the particular example in [34], on this remote server, a Python script featuring a pre-trained dense captioning model is utilised to produce captions for the image. Subsequently, these generated captions are transmitted back to the user's smartphone through the cloud system, where they are later converted into audio. For the human-robot interaction, when a human's image is captured by the camera of the robot, the image is sent to the server with pre-trained dense captioning to generate captions as useful instructions for the robot. Once the robot receives these instructions, it further processes it to finally take proper action for the user.

Existing dense captioning algorithms always adhere to two trajectories: Non-context dense captioning and context-guided dense captioning.

1) Non-context Dense Captioning: The pioneering framework proposed by [20] consists of a Region Proposal Network (RPN) based on Faster R-CNN, equipped with a bi-linear interpolation location module as an encoder and an LSTM as a decoder. Initially, all proposals are symbolized by the aligned features. These features are later fed into the RPN for binary foreground RoI detection. Identified foreground anchors are transformed into RoIs with corresponding features, with slight adjustments to bounding box coordinates. Finally, RoIs are captioned using an LSTM captioning model.

2) Dense Captioning With Context: To further enhance dense captioning performance, contextual knowledge was introduced to guide the task. [23] pioneered to incorporate context into dense captioning. In contrast to [20], this approach fused region features with global context features extracted from the entire image to generate descriptions. While the additional context improved the performance, it was observed that this form of contextual knowledge was too coarse to encode fine-grained context effectively.

In pursuit of capturing more fine-grained and detailed context, following endeavors were made. For instance, [24]

incorporated RoI neighbors and target RoI features into a non-local similarity graph to guide caption generation. Building on this, [19] recognized the potential of objects in images to provide valuable cues for locating RoIs, generating descriptions, and offering corresponding evidence through data statistics. Motivated by this revelation, authors pre-trained an offline object detector as guidance information for model training. Addressing the long-dependencies issue of LSTM and the equal weighting of RoIs in [19], [17] leveraged Transformer architecture to encode objects and regions. They also introduced a specialized loss module (ROCSU) to determine region-object correlation scores, resulting in significant performance improvements. Additionally, [18] identified the importance of textual context in cooperation with visual context during the captioning process, as well as the need for improved word training efficiency. To address these challenges, they introduced a Textual Context Module (TCM) within the Transformer decoder and a Dynamic Vocabulary Frequency Histogram (DVFH) to tackle these issues.

C. Transformer-based Object Detectors

Transformer-based architecture was first invented by [11] and many following variants focused on the refinement of the model structure. In particular, the Swin-Transformer scheme was proposed in [35] to gain a better representation of images. In addition, [36] introduced a Transformer in Transformer framework, taking advantage of smaller patches for more fine-grained features.

Recognising the Transformer's effectiveness in handling long sequential data, it found success in various computer vision tasks, including object detection. [37] took the initiative to introduce DETR, an end-to-end pipeline for object detection, which framed object detection as a set matching problem between queries and ground-truth bounding boxes. Subsequently, several upgraded versions, such as deformable DETR [38], were proposed to accelerate training and improve performance by focusing on local sampling points around a reference. Despite the achievements in object detection, the potential of Transformer-based models in addressing the challenging task of dense captioning remains underexplored. Drawing inspiration from the concise structure of traditional Faster R-CNN frameworks without artificially designed components, we integrated the DETR detector into the challenging dense captioning task.

III. Methodology

Our end-to-end dense captioning framework, shown in Fig. 3, is comprised of several key components. When presented with an image, our Transformer encoder utilizes attention mechanisms to encode the features extracted from the backbone. The RoI detection decoder, composed of two sequential cascade decoder layers, performs both RoI classification and regression, followed by a prediction head, in accordance with [37]. Our elaborately designed module, known as MSLD, employs a multi-scale supervision scheme in conjunction with a parallel decoder that takes cues from one of the two RoI detection decoders as its input. This innovative approach allows to generate discriminative multi-scale features that are instrumental for the captioning task. Furthermore, our MSLD leverages this multi-scale supervision scheme to provide a variety of feature queries, guiding feature learning for dense captioning across different object scales. To facilitate feature learning across diverse scales, we employ our MSLD module through a multi-scale extension, utilizing features from various layers. Consequently, the resulting multi-scale captioning features are harnessed to generate captions for regions of interest (RoIs).

In the upcoming sections, we will initially introduce our end-to-end dense captioning framework. Subsequently, we will delve into the deployment of our novel multi-scale language decoder (MSLD). Finally, we will provide detailed insights into model training and optimization.

A. Dense captioning framework via multi-scale transformer decoding (DCMSTRD)

1) Visual Feature Extraction: To capture rich features in a given image, in our DCMSTRD, we adopt a pre-trained ResNet-101 [39] backbone as in the work of [37] to extract the image features. To be specific, given an image $x \in R^{3\times H_0\times W_0}$, the ResNet-101 backbone, derives a lower-resolution activation map $f \in R^{C\times H\times W}$, $(C = 2048)$. To fit the sequential input of the following Transformer encoder, we collapse f to a feature map $F \in R^{HW\times C}$.

2) Transformer Encoder: The transformer encoder takes F as input. The choice of the number of Transformer encoder layers is determined empirically, taking into account two key factors. First, dense captioning involves two complex subtasks: RoI localization and RoI captioning. In similar compound tasks, such as dense video captioning in works like [40] and [41], as well as dense captioning in [18], it has been observed empirically that using two Transformer layers yields effective results. The second factor is the potential memory constraints. Compound tasks are often computationally demanding, and using too many Transformer layer stacks can lead to memory issues. To mitigate this, we opt for a Transformer encoder with two encoder layers. Each of these encoder layers follows a standard architecture, comprising a multi-head self-attention module and a feed-forward network (FFN). Additionally, positional encodings, as described in [42] and [43], are incorporated into the input of each attention layer. The outputs of these encoder layers are represented as E_1 and E_2, both of which are in the form of $R^{HW\times C}$.

3) RoI detection decoder: The RoI detection decoder, which also constitutes two decoder layers, takes E_2 and N RoI queries as inputs and N is remarkably larger than the typical number of RoIs in an image. Each decoder layer contains a standard self-attention layer and an encoder-decoder attention layer as [37]. In this way, the N RoI queries are transformed into the output features from each decoder layer, which are represented as D_1 and D_2 both belonging to the space $R^{N\times C}$. They are then independently decoded into box coordinates and class labels by two fully connected layers, resulting in N final predictions before binary set matching with the groundtruth.

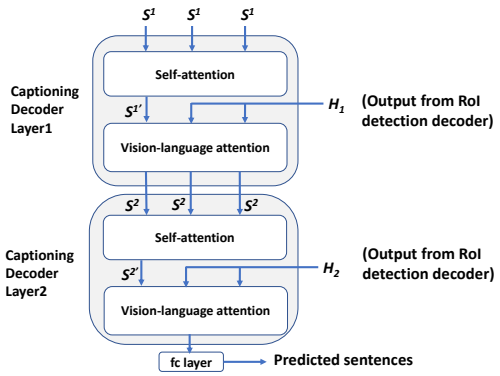


Fig. 4: The structure of our novel Multi-scale language decoder (MSLD).

4) Set matching: After obtaining the box coordinates and class labels from the features of each decoder layer D_1, D_2, our DCMSTRD matches predicted RoIs with the groundtruth, taking into account both RoI classification and box coordinates. Specifically, we adopt the Hungarian matching algorithm following [41] to assign the optimal bipartite matching results. The matching cost function is defined as follows:

$$Cost = \alpha_{cls}C_{cls} + \alpha_{giou}C_{giou} + \alpha_{RoI}C_{RoI}, \quad (1)$$

where C_{cls} represents the binary cross entropy loss of the predicted classification score and the groundtruth label, C_{giou} represents the generalized IoU [44] of the RoI and groundtruth coordinates, which are the center coordinates and the height and width relative to the image size as in [37]. C_{RoI} represents the l_1 loss between the predictions of the RoI coordinates and the groundtruth RoI coordinates. α_{cls}, α_{giou}, and $\alpha_{C_{RoI}}$ are hyperparameters set as 1.0, 2.0, 5.0 after [37]. These coefficient values are empirical. It considers the information from two aspects. Firstly, dense captioning is a compound task, consisting of two complex tasks: RoI localisation and RoI captioning. We found that many works on similar compound tasks empirically adopted the similar group of values for these coefficients. For example, [41] for dense video captioning, and [45] for pedestrian search. The second reason for this choice of numbers is that the RoI localisation is one of two sub-tasks, which are relatively more important than the RoI classification and therefore allocated a larger coefficient. After conducting the matching algorithm, the matched RoIs are selected, and the corresponding multi-scale features from D_1, $D_2 \in R^{N\times C}$, denoted as H_1, $H_2 \in R^{n\times C}$, where n represents the RoI number of the groundtruth in the given image. H_1 and H_2 also acts as the input of the multi-scale language decoder (MSLD) that will be introduced in the following section.

B. Multi-scale dense captioning decoder (MSLD)

Scale variation poses a major challenge in dense captioning as demonstrated in Fig. 2. To tackle this issue, we introduce an innovative and simple extension of our multi-scale language decoder (MSLD). The structural layout of this extension is depicted in Fig. 4. To extract multi-scale captioning features, it utilizes the features D_1 and $D_2 \in R^{N\times C}$, as introduced in Section III-A3, during the testing phase, and relies on word embeddings S as input during training. Comprising two layers to maintain consistency with the RoI detection decoder, each captioning decoder layer incorporates a self-attention layer for fine-tuning language features and a vision-language attention layer, which aids in learning multi-scale representations for the captioning task. The process of MSLD is outlined as follows:

$$SA(S^l_{\leq t}) = \begin{pmatrix} LN(MA(s^l_1, S^l, S^l), s^l_1) \\ \dots \\ LN(MA(s^l_t, S^l, S^l), s^l_t) \end{pmatrix};$$

$$VLA(S^l_{\leq t}) =$$

$$\begin{pmatrix} LN(MA((SA(S^l_{\leq t})_1), H_l, H_l), SA(S^l_{\leq t})_1) \\ \dots \\ LN(MA((SA(S^l_{\leq t})_t), H_l, H_l), SA(S^l_{\leq t})_t) \end{pmatrix}; \quad (2)$$

$$S^{l+1}_{\leq t} = LN(FFL(VLA(S^l_{\leq t})), VLA(S^l_{\leq t}));$$

$$p(w_{t+1}|S^L_{\leq t}) = soft\max(W_V S^L_{t+1}),$$

where $W_V \in R^{V_s\times d_{emb}}$ represents a matrix comprising word embeddings for the entire dictionary encompassing all words across the dataset. $s^0_i, i = 1...t$ stands for a series of corresponding word embeddings for the sentence with a dimension of d_{emb}. The variable l indicates the layer index and takes on values from the set 1,2. $S^l_{\leq t} = (s^l_1, ..., s^l_t)$ represents the predicted words up to time step $t+1$, with a triangular matrix masking out word information beyond this step. Here, SA denotes the self-attention layer, and MA corresponds to the multi-head attention, following the model proposed in [11]. LN stands for layer normalization, as described in [46]. VLP signifies the vision-language attention mechanism, which attends to language features up to step t, incorporating multi-scale representations H_l from the RoI detection decoder. FFL represents the feed-forward layer, following the formulation in [11], and $p(w_{t+1}|S^L_{\leq t})$ refers to the probability distribution over each word in the dictionary at time step $t+1$.

C. Training and Optimization Details

In this section, we will delve into the training and optimization specifics of our experiments. To ensure that both the localization of detected Regions of Interest (RoIs) and descriptive captions closely align with ground truth in an end-to-end fashion, we employ multiple loss components during each training iteration, as outlined below:

$$L = \lambda_{cls}L_{cls} + \lambda_{giou}L_{giou} + \lambda_{RoI}L_{RoI} + \lambda_{cap}L_{cap}, \quad (3)$$

where L_{cls} represents the binary cross entropy loss of the predicted RoI classification score and the groundtruth labels, L_{giou} represents the generalized IoU [44] of the matched RoI (excluding the unmatched background) bounding boxes and the groundtruth coordinates, in the form of the center coordinates and their height and width relative to the image size as in [37]. L_{RoI} represents the l_1 loss between the matched predictions of the RoI coordinates and groundtruth

RoI coordinates. L_{cap} is the cross entropy loss of $P = \{p(w_i|\theta), i \in [1, Sen_{max}]$, which is the probability distribution of descriptive sentences for RoIs in the RoI batch, and their groundtruth sentences word by word. θ represents all the trainable parameters of the whole system and Sen_{max} is the pre-set maximum word number in each sentence, and it is set to 10 for our experiments. λ_{cls}, λ_{giou}, $\lambda_{L_{RoI}}$ and λ_{cap} are balance coefficients, which are set as 1.0, 5.0, 5.0, 5.0.

IV. Results and Discussions

In this section, we report the results, discussion, and analysis of our experiments conducted on three publicly available datasets to assess the effectiveness of our proposed DCMSTRD algorithm.

A. Datasets and Evaluation Metrics

We employ two types of datasets, namely the Visual Genome dataset (VG) [47] and the VG-COCO dataset [19], for our evaluation benchmarks. This choice aligns with state-of-the-art methods [17]–[19], ensuring a fair comparison. We provide detailed descriptions of each dataset, along with the primary evaluation metrics.

1) Visual Genome (VG): VG is a large-scale dataset for dense captioning. There are 77,398 images in the training split and 5,000 images in the validation and test split, respectively. It has two widely used versions: VG V1.0, VG V1.2. The images of these are the same but the RoI groundtruth sentences are different. For consistency with the experimental settings in [17]–[19], [24], we also conduct our experiments on VG V1.0 and VG V1.2. The training, validation, and test splits are chosen similarly to [17]–[20], [24].

2) VG-COCO: As elaborated in [19], the RoI bounding boxes in 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 [48]. To obtain both fairer object bounding boxes and RoI bounding boxes for each image, following the configuration in [19], the intersection set of VG V1.2 and MS COCO is adopted in our paper, which is denoted as VG-COCO. VG-COCO is a smaller dataset than VG and there are 38,080 images for training, 2,489 images for validation and 2,476 for testing.

3) Evaluation Metrics: We also utilize the mean Average Precision (mAP) metric after [17]–[20], [24]. mAP assesses the precision of both localisations and RoI captions whilst the mAP in object detection considers the object localisation and the accuracy of the classification. Following the threshold settings in [17]–[20], [24], average precision is calculated under different combinations of IoU thresholds (0.3, 0.4, 0.5, 0.6, 0.7) to evaluate the predicted RoI locations and Meteor [48] thresholds (0, 0.05, 0.10, 0.15, 0.20, 0.25) to assess the similarity between predicted RoI captions the ground truth sentences. With each group of thresholds (30 groups in aggregate), the Average Precision (AP) can be calculated. Finally, the mean value of these APs is the mAP score of all 30 threshold combinations.

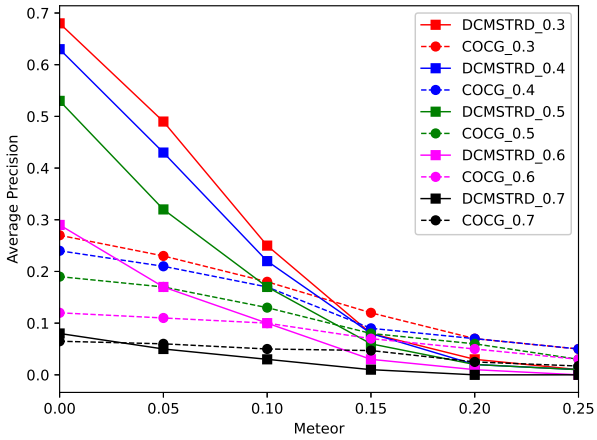


Fig. 5: Average precision of DCMSTRD and COCG methods under different threshold combinations of Meteor scores and IoU on the VG-COCO dataset.

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

Method	mAP(%)
FCLN [20]	4.23
JIVC [23]	7.85
Max Pooling [19]	7.86
COCD [19]	7.92
COCG [19]	8.90
ImgG [19]	7.81
COCG-LocSiz [19]	8.76
COCG&GT [19]	9.79
TDC+ROCSU [17]	11.58
ETDC(VGG16) [18]	12.28
ETDC+TCM+DVFH [18]	14.30
DCMSTRD(Ours)	**16.10**

B. Implementation Details

Our DCMSTRD is trained on a single NVIDIA GTX 2080 Ti GPU with a memory of 11GB using AdamW optimizer as [37]. The image batch size is set to 1, the epoch is set to 60. The initial learning rate is 10^{-5} and the backbone learning rate is 10^{-6}. The learning rate drop factor is 0.1 at epoch 50. The momentum factor is set to 0.9, and weight decay is 10^{-4}. The RoI query number N is set to 1,500 and the multi-head number in Eq. 2 is 8.

C. Quantitative Results and Analysis

In this section, we first display quantitative results and discussions on three publicly available datasets: VG-COCO, VG V1.0 and VG1.2 respectively. Subsequently, we quantify the performance of our proposed DCMSTRD and MSLD through ablation studies.

1) Experimental results, discussions, and analysis on VG-COCO Dataset: In the evaluation of the VG-COCO dataset, we conducted a comparative analysis of our DCMSTRD architecture alongside other baseline methods, as presented in Table I. The results reveal a notable difference in mAP, with DCMSTRD achieving a significant improvement of 16.10%. Com-

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 [20]	5.39	5.16
JIVC [23]	9.31	9.96
ImgG [19]	9.25	9.68
COCD [19]	9.36	9.75
COCG [19]	9.82	10.39
CAG-Net [24]	10.51	—
ETDC(VGG16)	11.31	10.60
TDC+ROCSU [17]	11.49	11.90
ETDC+TCM+DVFH [18]	13.24	12.60
DCMSTRD	**13.63**	**13.44**

pared to the state-of-the-art ETDC+TCM+DVFH approach in [18], DCMSTRD demonstrates a remarkable gain of 1.8 in terms of mAP. Additionally, when compared to the leading LSTM method, COCG, DCMSTRD exhibits a substantial mAP increase of 80.9%. Our method's performance superiority over other approaches is even more pronounced, with DCMSTRD achieving a mAP that is over three times higher than the initial FCLN method. Furthermore, when compared to other comparative methods such as JIVC, Max Pooling, and COCD, DCMSTRD achieves an absolute mAP improvement of approximately 8. Notably, even when combining ground truth localization of each RoI with the state-of-the-art method COCG in COCG&GT, DCMSTRD still outperforms it with a remarkable 65.45% increase in mAP. These results underscore the superiority of DCMSTRD, attributable to the elimination of artificially designed components, enhanced interactions between sub-tasks, and the multi-scale feature representations offered by our innovative MSLD module.

2) Experimental results, discussions, and analysis on VG V1.0 : We also evaluated DCMSTRD on the VG V1.0 dataset, and the results are presented in the second column of Table II. DCMSTRD achieved a notable mAP score of 13.63, surpassing all previous methods by a substantial margin on this dataset. Specifically, our method outperformed ETDC+TCM+DVFH [18], TDC+ROCSU [17], and the COCG method [19] by margins of 0.39, 2.14, and 3.81, respectively. Furthermore, when compared to CAG-Net in [9], DCMSTRD exhibites a significant improvement of 3.12 mAP. This improvement can largely be attributed to the discriminative multi-scale features learned by the MSLD module within DCMSTRD. This module integrates scale information from generated captions, thus enhancing precision. Moreover, DCMSTRD's advantage stems from its end-to-end learning approach, which avoids the need for manually set thresholds seen in artificially designed components like Faster R-CNN. This approach contributed to the superior performance mentioned earlier. It's worth noting that the mAP increase against state-of-the-art methods on VG V1.0 is smaller compared to VG-COCO. This discrepancy is likely due to VG V1.0 being more than twice the size of VG-COCO, containing more RoIs with complex scenes and captions. Consequently, describing these RoIs becomes significantly more challenging.

3) Experimental results, discussions, and analysis on VG V1.2: We conducted tests of our DCMSTRD approach on the VG V1.2 dataset, and the resulting mAP scores are presented in the third column of Table II. DCMSTRD achieves notable relative mAP improvements, surpassing the ETDC+TCM+DVFH, TDC+ROCSU (11.90), and COCG (10.39) methods by margins of 0.84, 1.54, and 3.05, respectively, with an overall mAP of 13.44.

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

DCMSTRD	MSLD	mAP(%)
✗(Faster R-CNN)	✓	12.56
✓	✗	13.02
✓	concat	14.97
✓	✓	**16.10**

Remarkably, the mAP achieved by our DCMSTRD framework exceeded the mAP of the FCLN method by more than twice, underscoring the effectiveness of our proposed DCMSTRD approach and MSLD architecture. Notably, on VG V1.2, as with VG V1.0, the performance gap between our method and other prior approaches is narrower compared to VG-COCO. This could be attributed to the similar data distributions in VG V1.0 and VG V1.2 (the same image set with slightly different corresponding captions), resulting in more RoIs with complex visual scenes and corresponding captions, which pose greater challenges.

4) Comparison of average precision under different RoI detection and description thresholds: Fig. 5 shows the comparison of average precision between our proposed DCMSTRD and the COCG method in [19] as it is the only publicly available that is comparable. It is easily seen that overall, our proposed DCMSTRD outperforms the COCG method. In particular, DCMSTRD performs much better than COCG under low thresholds of both detection and language. This is due to two aspects: The general DCMSTRD applies RoI queries, which can better learn the distribution of the locations of all the RoIs thus providing a global optimization for the RoI detection problem. Hence it is straightforward to generate qualified RoIs against groundtruth RoI bounding boxes under lower thresholds. Furthermore, our MSLD module successfully learns the multi-scale features and enables the ability to recognize objects in different scales thus creating more good-quality descriptions under lower language thresholds. However, under higher thresholds (IoU$\geq$0.6 and Meteor$\geq$0.15), COCG method slightly outperforms DCMSTRD because in COCG, a pre-trained object detector was deployed and thus providing more valuable priors both for RoI locations and RoI captioning. These extra clues help the model create some relatively better RoI locations and captions under higher thresholds.

5) Ablation Studies: To validate the effectiveness of our proposed DCMSTRD method and the MSLD module, we also perform a variety of ablation studies. First of all, to validate the advantage of our DCMSTRD framework, we preserve the MSLD module but replace the RoI detection decoder with a Faster R-CNN framework as [18]. It should be noted that to carry out a fair comparison, we also delete the prior knowledge of the pre-trained objects in the visual encoder in [18]. This method is denoted as Faster R-CNN with multi-scale language decoder (FRMSLD) as shown in

the first row of Table III. It can be observed that without the general DCMSTRD architecture, the FRMSLD fails to choose the suitable parameters of the artificially designed components inside for all the images across the whole dataset, and therefore, the mAP drops to 12.56, which is inferior to the proposed DCMSTRD+MSLD (16.10).

Moreover, we also develop a degraded model which only maintains DCMSTRD but removes the MSLD module, denoted as DCMSTRD-MSLD as shown in the second row of Table III to verify the impact of the MSLD module. To be specific, the captioning decoder of DCMSTRD-MSLD is implemented as follows:

$$
\begin{aligned}
&SA(S^l_{\leq t}) = \begin{pmatrix} LN(MA(s^l_1, S^l, S^l), s^l_1) \\ ... \\ LN(MA(s^l_t, S^l, S^l), s^l_t) \end{pmatrix}; \\
&VLA(S^l_{\leq t}) = \\
&\begin{pmatrix} LN(MA((SA(S^l_{\leq t})_1), H_2, H_2), SA(S^l_{\leq t})_1) \\ ... \\ LN(MA((SA(S^l_{\leq t})_t), H_2, H_2), SA(S^l_{\leq t})_t) \end{pmatrix}; \\
&S^{l+1}_{\leq t} = LN(FFL(VLA(S^l_{\leq t})), VLA(S^l_{\leq t})); \\
&p(w_{t+1}|S^L_{\leq t}) = soft\max(W_V S^L_{t+1}),
\end{aligned}
\tag{4}
$$

The notation is the same as Eq. 2. Note that in DCMSTRD-MSLD, we only adopted the visual hidden states of the last RoI detection decoder H_2 into the captioning decoder without multi-scale dense captioning decoding anymore. The difference between DCMSTRD method and DCMSTRD-MSLD method is the existence of MSLD module (DCMSTRD is with MSLD module while in DCMSTRD-MSLD it is removed). The MSLD module fuses the multi-scale features while DCMSTRD only takes the hidden states from the second layer of RoI detection decoder. It can be seen that without MSLD module, the performance is slightly better than FRMSLD by 0.46. Nonetheless, due to the missing MSLD, the machine is incapable of learning multi-scale invariant features. Hence, it suffers from the generation of poor captions for the RoIs of the same class but with different scales or even fails to detect a part of multi-scale RoIs. As a consequence, the mAP of DCMSTRD-MSLD is 3.08 less than the DCMSTRD method, standing at 13.02.

To further validate the effectiveness of the proposed MSLD module, we also detach MSLD as DCMSTRD-MSLD but input the concatenated output of RoI detection decoder H_1 and H_2 into the captioning decoder, denoted as DCMSTRD (concat). Its deployment is as follows:

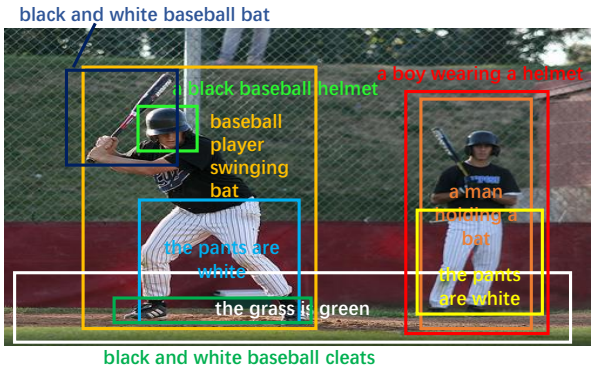
$$
\begin{aligned}
&SA(S^l_{\leq t}) = \begin{pmatrix} LN(MA(s^l_1, S^l, S^l), s^l_1) \\ ... \\ LN(MA(s^l_t, S^l, S^l), s^l_t) \end{pmatrix}; \\
&VLA(S^l_{\leq t}) = \\
&\begin{pmatrix} LN(MA((SA(S^l_{\leq t})_1), H, H), SA(S^l_{\leq t})_1) \\ ... \\ LN(MA((SA(S^l_{\leq t})_t), H, H), SA(S^l_{\leq t})_t) \end{pmatrix}; \\
&S^{l+1}_{\leq t} = LN(FFL(VLA(S^l_{\leq t})), VLA(S^l_{\leq t})); \\
&p(w_{t+1}|S^L_{\leq t}) = soft\max(W_V S^L_{t+1})
\end{aligned}
\tag{5}
$$

where H is the concatenation of the multi-scale features H_1 and H_2, and other notations are as Eq. 2. According to the experimental results in the last second row in Table III, when concatenated RoI detection hidden states H is used to replace the MSLD module, the mAP performance decreases to 14.97 but is still lesser than our proposed DCMSTRD by 1.13. This is because in DCMSTRD (concat), even if the multi-scale features are concatenated as input, the input for the two sub-layers of the captioning decoder is consistent and therefore fails to assign different tasks of tuning multi-scale features to different captioning decoder layers, which hinders the whole process of multi-scale feature tuning.

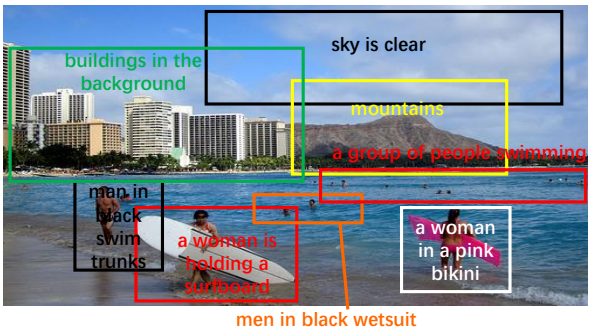
D. Qualitative Results and Discussions

The qualitative results and discussions are also exemplified in this section to assess our DCMSTRD method. In the first subsection, we show a couple of instances from the VG-COCO and VG V1.2 datasets via showing all RoIs and their generated captions from our DCMSTRD framework. Following these instances, we show the qualitative results of ablation studies of our proposed DCMSTRD method and the DCMSTRD-MSLD method. Lastly, we perform a comparison between DCMSTRD, the state-of-the-art COCG method, and the provided ground truth.

1) Visual examples of dense captioning by DCMSTRD method: In Fig. 6, we present two illustrative examples of dense captioning results obtained using our DCMSTRD method applied to two distinct datasets: VG-COCO and VG V1.2. The first instance showcased in Fig. 6a originates from the VG-COCO dataset, while the second example in Fig. 6b is drawn from the VG V1.2 dataset. These visual samples vividly exemplify the high-quality RoI detection and caption generation achieved by our proposed DCMSTRD method. First and foremost, our DCMSTRD model excels in generating descriptions for RoIs with impeccable grammatical structure. The majority of the generated sentences exhibit a high degree of accuracy and adhere to proper English grammar conventions. This proficiency can be primarily attributed to the inherent qualities of DCMSTRD, as it eliminates the need for artificially designed components and associated thresholds. Furthermore, DCMSTRD undertakes both sub-tasks concurrently, following an end-to-end approach, thereby facilitating the discovery of optimal solutions across the entire dataset.



(a) The example of dense captioning results achieved by DCMSTRD method from VG-COCO.



(b) The example of dense captioning results achieved by DCMSTRD method from VG V1.2.

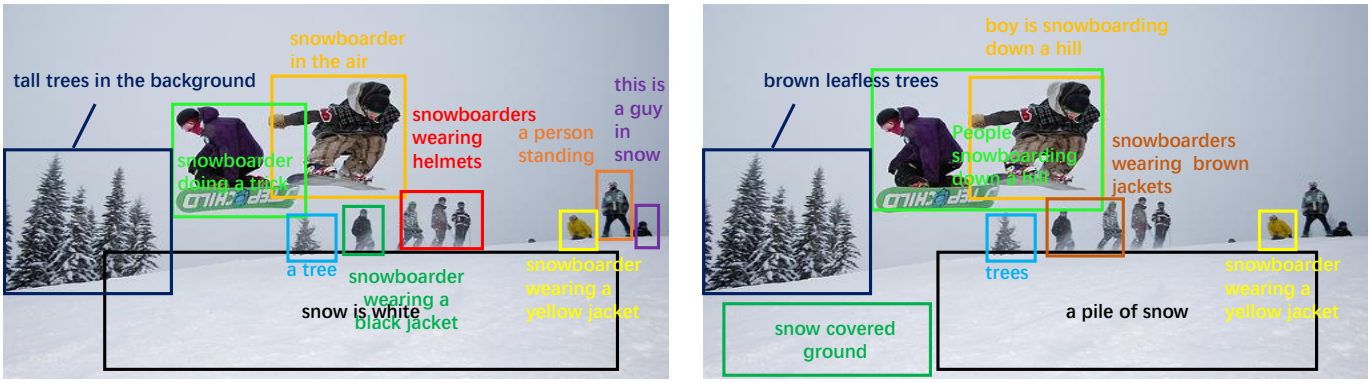
Fig. 6: Examples of dense captioning results achieved by our DCMSTRD framework.

Moreover, our proposed DCMSTRD method demonstrates competence in detecting RoIs of varying scales. e.g., in the first example in Fig. 6a, it adeptly identifies and characterizes RoIs containing two baseball players, each at a different scale. It also successfully distinguishes between distinct states or actions of the objects within these RoIs, generating relevant descriptions such as 'baseball player swinging bat' for the first baseball player RoI and 'a man holding a bat' for the second. In Fig. 6b, we further observe the proficiency of DCMSTRD in identifying RoIs encompassing relatively large individuals, depicted within the black, red, and white RoIs on the beach. These are accompanied by precise and descriptive captions. Impressively, the method extends its capabilities to detect individuals at a considerably smaller scale, as evidenced by the red RoI in the sea, accompanied by the apt description 'a group of people swimming' These remarkable advantages are attributable to our innovative MSLD module, which equips the system with the capability to recognize and describe objects effectively across a spectrum of scales.

2) Ablation studies: In this section, we undertake an in-depth evaluation of the experimental results derived from our DCMSTRD and MSLD components, individually dissecting the significance of each contribution. Specifically, we present dense captioning results for both DCMSTRD and the DCMSTRD-MSLD variants within the same image from VG V1.0, as shown in Fig. 7. Overall, the DCMSTRD approach exhibits a heightened focus on delivering precise and detailed descriptions, particularly when it comes to people depicted at various scales within the given image. Firstly, the light green boxes associated with both methods highlight that DCMSTRD excels in providing specific descriptions of individual actions, such as 'snowboarder doing a trick', while DCMSTRD-MSLD tends to produce relatively generic RoI captions like 'people snowboarding down a hill'. Moreover, DCMSTRD demonstrates a greater propensity to detect each RoI featuring a person independently, whereas DCMSTRD-MSLD consistently describes groups of people together. Furthermore, it is evident that our proposed DCMSTRD method proficiently identifies all individuals on the snow slope, even those at a minute scale, as exemplified by the person in the purple bounding box with the caption 'this is a guy in snow'. Conversely, DCMSTRD-MSLD tends to overlook most of these individuals. These findings can be attributed to two key factors: firstly, DCMSTRD eliminates the need for artificially designed components and their associated thresholds, thereby enhancing the overall RoI detection process and performance. Secondly, the MSLD component augments the learning of multi-scale features, consequently bolstering the model's ability to discriminate between individuals or objects at diverse scales and improving its comprehension of specific human actions or states. As a result of this, the model with MSLD can strengthen its ability to discriminate people or objects at various scales and have a better understanding of the specific motion taken by a person or the state of a person.

3) Convergence Process of Training Loss: To better display the training process of our proposed model, we show the loss (objective) function value of our proposed model during training in Fig. 8. We can observe that the loss function value plummets during the initial several epochs, followed by a gradual but slower decline until the 50th epoch where the learning rate drops. The effect of this drop is considerable, and the loss value starts to go down again due to the smaller and careful pace during the optimisation. Eventually, the overall loss keeps a trend of convergence at epoch 60, standing at the value of around 2.25.

4) Comparative results with COCG method and ground truth: Fig. 9 presents comparative qualitative results among our DCMSTRD method, the state-of-the-art COCG method, and the ground truth, serving as a straightforward means to visually assess their performance. It is obvious that the proposed DCMSTRD method achieves better performance for both RoI detection and RoI description due to higher IoUs and Meteor scores shown in the figure. It is noticeable that DCMSTRD is expected to significantly surpass the COCG method in Meteor language score performance. Our proposed DCMSTRD method and MSLD module both contribute to this better performance. On one hand, the DCMSTRD pipeline can skip the sub-optimal parameter settings of artificially designed components in prior works, thus leading to an optimal detection result. For instance, in Fig. 9a, our DCMSTRD



(a) The visualization results of DCMSTRD method.

(b) The visualization results of DCMSTRD-MSLD method.

Fig. 7: The comparative qualitative results of the DCMSTRD method and the model that removes the MSLD module.

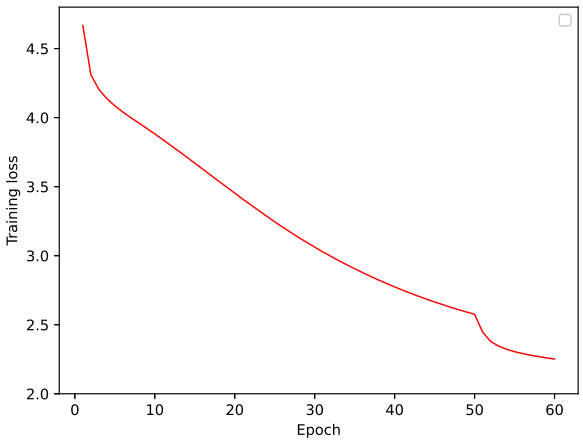


Fig. 8: Loss (objective) function value of training.

method can cover the arm of the man whilst COCG method only contains the body part due to the sub-optimal thresholds predefined in the artificially designed modules. Furthermore, our MSLD can benefit representative feature learning, especially for images with multi-scale information. In Fig. 9b, our proposed MSLD framework can successfully recognize multi-scale information of different horses and people, thus generating satisfactory captions with a Meteor score of 0.26. However, without taking multi-scale information into account, COCG fails to learn the different shapes and the relations between a group of people and horses and is impotent to produce correct captions.

V. Conclusion

In this paper, a novel trainable dense captioning architecture, termed end-to-end dense captioning framework via multi-scale transformer decoding (DCMSTRD) is introduced. In particular, DCMSTRD replaces the artificially designed modules and reschedules the dense captioning task as a set prediction and matching problem. Furthermore, the learning of multi-scale representations is very important to dense captioning. To this end, we also proposed a novel module, named multi-scale language decoder (MSLD). We assessed our innovative approach on three publicly available datasets and the results show that our method surpassed state-of-the-art methods by a wide margin in terms of mean Average Precision. Owing to its application portability, in our future work, we will look at applying our model to different close tasks such as image segmentation [4], event detection [49], object detection [50], [51], pedestrian detection [52], [53], pedestrian attribute recognition [54], person search [45], [55], 3D model retrieval [56], [57], zero-shot learning [58], and magnetic resonance imaging [59] though there should be some adjustment on DCMSTRD and MSLD modules according to adjust to the requirements of each specific task.

References

[1] 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.

[2] T. Qian, J. Chen, S. Chen, B. Wu, and Y.-G. Jiang, "Scene graph refinement network for visual question answering," *IEEE Transactions on Multimedia*, vol. 25, pp. 3950–3961, 2023.

[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] 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.

[5] J. Yang, W. Liu, J. Yuan, and T. Mei, "Hierarchical soft quantization for skeleton-based human action recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 883–898, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

(a) (b)

Fig. 9: Qualitative results of baseline (COCG) and our proposed method (DCMSTRD). 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 DCMSTRD approach (Best viewed in color).

[10] W. Jiang, W. Wang, and H. Hu, "Bi-directional co-attention network for image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 4, pp. 1–20, 2021.

[11] 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.

[12] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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*, pp. 1–12, 2022.

[18] Z. Shao, J. Han, K. Debattista, and Y. Pang, "Textual context-aware dense captioning with diverse words," *IEEE Transactions on Multimedia*, pp. 1–15, 2023.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] A.-A. Liu, H. Tian, N. Xu, W. Nie, Y. Zhang, and M. Kankanhalli, "Toward region-aware attention learning for scene graph generation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7655–7666, 2022.

[26] 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.

[27] N. Xu, A.-A. Liu, Y. Wong, W. Nie, Y. Su, and M. Kankanhalli, "Scene graph inference via multi-scale context modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1031–1041, 2020.

[28] 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.

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

[30] 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.

[31] 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.

[32] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[33] 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.

[34] N. A. Giudice and G. E. Legge, "Blind navigation and the role of technology," *The engineering handbook of smart technology for aging, disability, and independence*, pp. 479–500, 2008.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[36] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[39] 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.

[40] 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.

[41] 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.

[42] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[43] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[44] 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/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] A. Gao, Y. Pang, J. Nie, Z. Shao, J. Cao, Y. Guo, and X. Li, "Esgn: Efficient stereo geometry network for fast 3d object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[51] Li, Zhihui and Xu, Pengfei and Chang, Xiaojun and Yang, Luyao and Zhang, Yuanyuan and Yao, Lina and Chen, Xiaojiang, "When object detection meets knowledge distillation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[52] F. Chu, J. Cao, Z. Shao, and Y. Pang, "Illumination-guided transformer-based network for multispectral pedestrian detection," in *CAAI International Conference on Artificial Intelligence*. Springer, 2022, pp. 343–355.

[53] N. Chen, J. Xie, J. Nie, J. Cao, Z. Shao, and Y. Pang, "Attentive alignment network for multispectral pedestrian detection," in *Proceedings of the 31st ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2023, pp. 3787–3795.

[54] Z. Ji, Z. Hu, Y. Wang, Z. Shao, and Y. Pang, "Reinforced pedestrian attribute recognition with group optimization reward," *Image and Vision Computing*, vol. 128, p. 104585, 2022.

[55] J. Wang, Y. Pang, J. Cao, H. Sun, Z. Shao, and X. Li, "Deep intra-image contrastive learning for weakly supervised one-step person search," *Pattern Recognition*, vol. 147, p. 110047, 2024.

[56] D. Song, Y. Yang, W. Li, Z. Shao, W. Nie, X. Li, and A.-A. Liu, "Adaptive semantic transfer network for unsupervised 2d image-based 3d model retrieval," *Computer Vision and Image Understanding*, vol. 238, p. 103858, 2024.

[57] L. Z. Jiacheng Chang and Z. Shao, "View-target relation-guided unsupervised 2d image-based 3d model retrieval via transformer," *Multimedia Systems*, 2023.

[58] Liu, Zhe and Li, Yun and Yao, Lina and Chang, Xiaojun and Fang, Wei and Wu, Xiaojun and El Saddik, Abdulmotaleb, "Simple Primitives with Feasibility-and Contextuality-Dependence for Open-World Compositional Zero-shot Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[59] 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 the 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).