

CDKM: Common and Distinct Knowledge Mining Network with Content Interaction for Dense Captioning

Hongyu Deng, Yushan Xie, Qi Wang*, Member, IEEE, Jianjun Wang, Weijian Ruan, Wu Liu, Senior Member
Yong-Jin Liu, Senior Member, IEEE

Abstract—The dense captioning task aims at detecting multiple salient regions of an image and describing them separately in natural language. Although significant advancements in the field of dense captioning have been made, there are still some limitations to existing methods in recent years. On the one hand, most dense captioning methods lack strong target detection capabilities and struggle to cover all relevant content when dealing with target-intensive images. On the other hand, current transformer-based methods are powerful but neglect the acquisition and utilization of contextual information, hindering the visual understanding of local areas. To address these issues, we propose a common and distinct knowledge-mining network with content interaction for the task of dense captioning. Our network has a knowledge mining mechanism that improves the detection of salient targets by capturing common and distinct knowledge from multi-scale features. We further propose a content interaction module that combines region features into a unique context based on their correlation. Our experiments on various benchmarks have shown that the proposed method outperforms the current state-of-the-art methods. The code is open at: <https://github.com/GZU-SAMLab/CDKM/>

Index Terms—Few-shot learning, Object detection, Multi-scale projection, Contrastive learning.

I. INTRODUCTION

DENSE captioning, as a crucial task, aims to solve the real-world problem of multimodal data understanding and application. Specifically, the purpose of dense captioning is to detect multiple salient regions of an input image and describe their content in natural language. The goals of this task are quite similar to those of open-set object detection [5], [13], [34], [37], [41] and image captioning [14], [31], [33], [36], [39]. On the one hand, compared to open-set object detection, dense captioning needs to describe Regions of Interest (RoIs) using relatively long language sequences rather

His research was supported by the National Natural Science Foundation of China (No. 62162008, 62332019), Beijing Nova Program (NO. 20220484063), Guizhou Provincial Science and Technology Projects (CXTD[2023]027, ZK[2022]-108), Guizhou Province Overseas Talents Innovation and Entrepreneurship Selective Funding (2022-04). Guiyang Guian Science and Technology Talent Training Project ([2024] 2-15). *: Qi Wang is the corresponding authors. (qiwang@gzu.edu.cn)

H. Deng, Q. Wang, Y. Xie, J. Wang are with the State Key Laboratory Of Public Big Data, Text Computing & Cognitive Intelligence Engineering Research Center of National Education Ministry, Guizhou University, Guiyang 550025, China.

Q. Wang and Y. Liu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

W. Liu, is with JD Explore Academy, JD.com, Beijing 100101, China.

W. Ruan is with the Shenzhen Institute of Advanced Technology Chinese Academy of Sciences, Shenzhen 518055, China

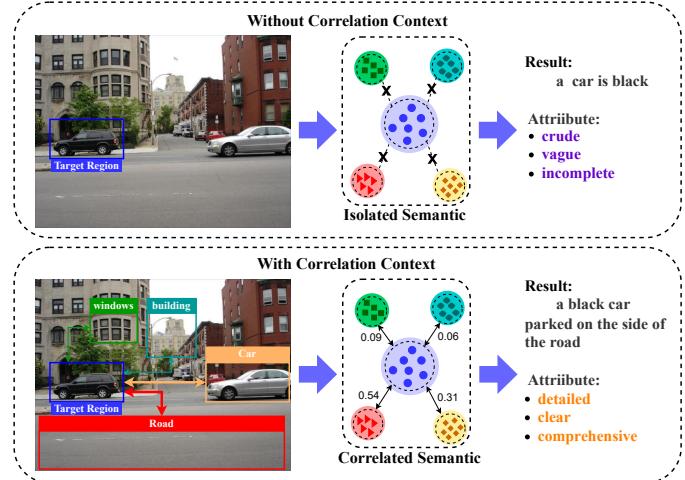


Fig. 1: Comparison of different captioning methods. The below method performs the region interaction and obtains context to assist captioning, while the above method does not. Apparently, context prompts captions to be more detailed, clear, and comprehensive.

than several words. On the other hand, dense captioning focuses on different local areas of the image as opposed to image captioning, thus conveying more details for the users. Therefore, dense captioning methods require stronger image understanding and language organization abilities. This technique has enabled computers to better understand and interpret visual information and thus play a greater role in a variety of applications, including navigation [8], human-computer interaction [22], video captioning [1], [38] and image-text retrieval [30]. All these applications show the strong potential of dense captioning and the prospect of wide application.

Dense captioning is a challenging task that has achieved great success in recent years. The dense captioning task is originally proposed by Johnson et al. [11], who also presents the first network for this task, called FCLN. Johnson et al. propose an excellent idea for the dense captioning task, containing several steps of feature extraction, region proposal, and region caption generation, which gradually evolves into a basic paradigm for dense captioning. However, these similar approaches inevitably suffer from the aperture problem, which makes it difficult to generate accurate descriptions. Some

studies have tried to improve the performance of dense captioning by introducing contextual information. Yang et al. [32] directly treat the global RoI pooling feature as a kind of visual context to guide a more accurate description, achieving a great improvement in performance. Li et al. [15] employ RPN and Faster R-CNN to obtain RoI features and salient target features from images, and fuse two kinds of visual features by LSTM to obtain more efficient contextual information. In addition, recent works have also introduced transformer structure into the research of dense captioning. For example, Shao et al. [24] proposed an end-to-end transformer-based framework that considers the relationship between objects and regions, as well as the confidence scores of detected objects, to measure the importance of each region. Wu et al. [28] and Long et al. [21] both have proposed a powerful transformer-based framework to unify dense captioning and open-set object detection, which achieves highly competitive performance on both tasks. Recently, Shao et al. [23] have proposed a novel framework that extracts textual context from images and dynamically diversifies the vocabulary bank during the captioning process, thus making the generated captions more diverse. Overall, the introduction of transformer structure and the use of contextual information have improved the performance of dense captioning, making it a challenging but promising area of research.

However, despite the efforts made by existing methods, there are still some drawbacks that must be addressed. First, many of these methods struggle to describe complex images effectively, as they often rely on relatively basic object detection modules to locate salient regions. These modules typically operate at a single scale, resulting in proposals that are scattered across multiple regions without providing critical information. Second, current transformer-based methods still suffer from the aperture problem, which limits their ability to capture the internal relationships between different regions. For example, GRiT, a method that unifies open-set object detection with dense captioning, treats each region as an independent visual element. This leads to fragmentation of the connections among different areas when performing the dense captioning task. Intuitively, the interaction among regions harbors a semantic link between the local and global aspects of an image, which can be utilized as context to alleviate the aperture problem, as demonstrated in Figure 1. We aim to capture and utilize these interactions among regions to enhance the visual understanding of the model and thus contribute to detailed, clear, and comprehensive captions.

Motivated by the above views, we propose a transformer-based dense captioning framework. First, we propose a common and distinct knowledge-mining network that utilizes multi-scale features and implicit knowledge to address the first drawback. We acquire the multi-scale features from images by the ViT [4] and FPN [17]. Then, the knowledge-mining network learns the shared and distinct knowledge hidden within these multi-scale features, enabling it to accurately capture targets of varying sizes. Second, to address the second drawback, we introduce a content interaction module for visual contextual information acquisition. This module employs a content interaction mechanism to compute the

correlation between different salient regions within the image and generate context tokens for each region. These context tokens are concatenated with regional features to enhance the model's visual understanding and improve its ability to detect and understand complex visual scenes. Overall, with the help of the knowledge-mining network and content interaction module, our proposed framework addresses the challenges associated with dense captioning. Our results demonstrate the effectiveness of our proposed approach in generating high-quality and coherent captions for various visual scenes.

In short, our contributions are summarized as follows:

- We design a knowledge mining network based on implicit information and content interaction that can identify significant regions more comprehensively and accurately in target-dense situations.
- We propose to capture the correlation between different salient regions and leverage it as contextual information to enhance the model's visual understanding.
- Extensive experiments conducted on multiple datasets demonstrate that our method outperforms the current state-of-the-art methods.

This paper is organized as follows. Section II introduces the related works. Section III presents the proposed model in detail. Section IV conducts the experiments with both qualitative and quantitative analysis. Section V summarizes our work.

II. RELATED WORKS

A. Open-set Object Detection

Open-set object detection [6], [20] is a challenging vision task intended to detect and classify objects in an image or video, even if they belong to an unknown class not seen during training. Since this task also requires the model to be flexible in its perception and discrimination of targets, a number of worthy works are illuminating for the research of dense captioning. Currently, some works attempt to utilize large-scale vision-language pretraining to empower object detection models to recognize unseen and unknown categories. For example, Zareian et al. [37] propose a novel framework called open vocabulary object detection, which enables accurate localization of unknown objects. Gao et al. [5] utilize the localization capabilities of pre-trained vision-language models to generate pseudo-bounding box labels, thus providing more abundant base-class data for the object detector. Yao et al. [34] propose a parallel visual concept pre-training method, DetCLIP, for open-world detection by exploiting knowledge enrichment in a designed concept dictionary. These approaches that leverage the internal knowledge of pre-trained models to improve visual understanding have motivated some of the work on dense captioning [21], [28].

B. Image Captioning

Image captioning [9], [25], [26] is a rather basic and essential multimodal task, which requires the models to realize a whole scene of an image and to generate a summarized description of the content in natural language. Due to the large

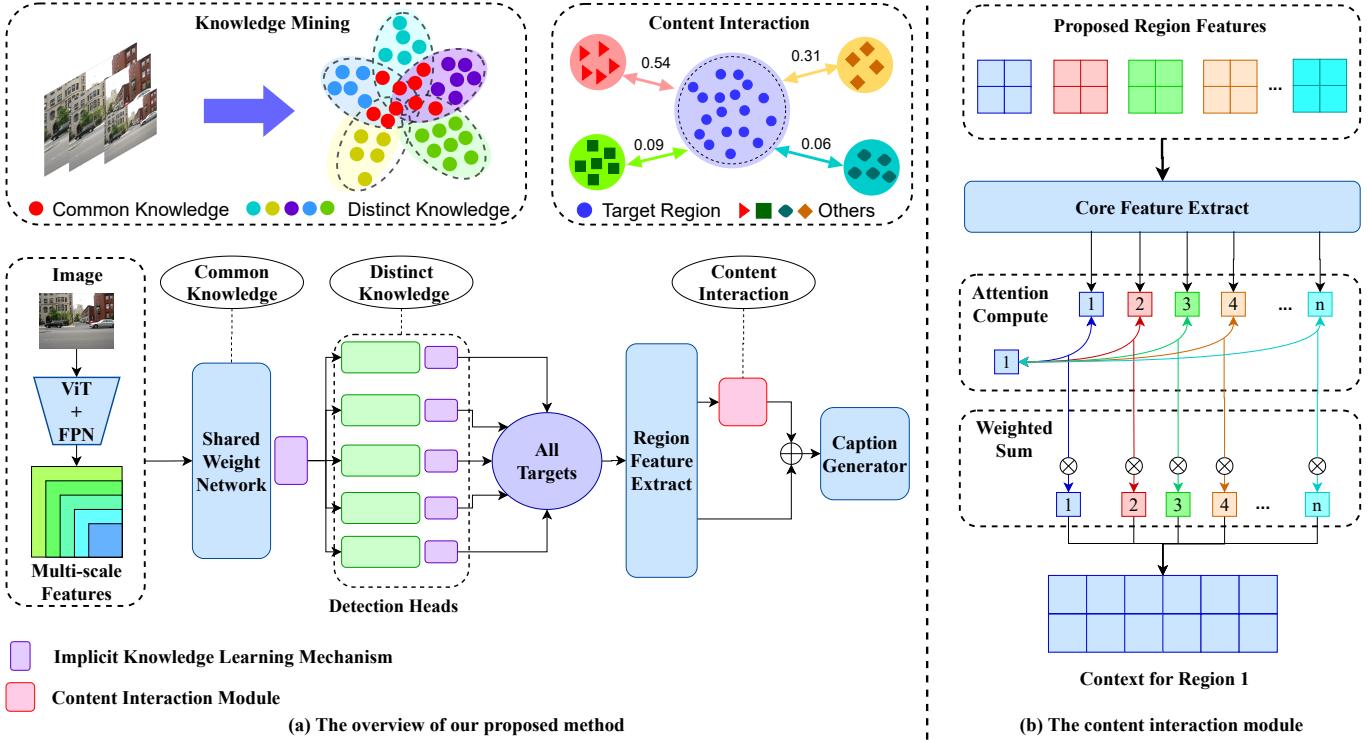


Fig. 2: The framework of the proposed method. (a) The overview of our proposed method. Common and distinct knowledge of multi-scale features is summarized in the knowledge mining network. The content interaction module generates correlation context for the target region based on semantic relatedness between regions. (b) The calculation process of the content interaction module. The content interaction module facilitates the model's understanding of the local areas by calculating the correlation of a region with other salient regions, using the features of the other regions to constitute context exclusive to that region.

commonality between image captioning and dense captioning, approaches specializing in these two tasks can often build on and translate into each other. Extracting salient image regions with object detection networks has been one of the mainstream practices in image captioning methods over recent years, which can be traced back as far as the work of Anderson et al. [2]. The BUTD proposed by Anderson et al. adopts Faster R-CNN to localize salient objects in images and extract the corresponding region features. Subsequently, by implementing top-down attention on the region features, contextual information is obtained to predict the attention distribution over the image regions. This method greatly inspires subsequent works and has a far-reaching impact. By leveraging the attention result and the current context to obtain attended information, Huang et al. [10] extend the traditional attention mechanism in the encoder and decoder, thus determining the relevance between the attention result and the query. Yu et al. [36] apply dual attention to pyramid image feature maps to fully explore the visual-semantic correlations, which contributes to captions with high-quality. In the text-based image captioning (TextCap) task, Wang et al. [27] propose a collaborative attention module, which models the association between image regions and OCR token regions to compute two kinds of contextual information to improve image understanding. The above methods all exploit the internal association between salient regions to obtain context and thus improve the models, which greatly inspires our work.

C. Dense Captioning

Dense captioning aims to localize salient regions of an image and describe the content of each region using natural language, providing users with richer details compared to image captioning. Most current dense captioning methods are implemented based on CNN and RNN. The earliest method is the FCLN proposed by Johnson et al. [11], which employs a CNN to localize salient regions in an image and then uses an RNN to generate corresponding captions for the region features. However, FCLN suffers from the aperture problem [35]. Specifically, when lacking sufficient contextual information, the model may fail to realize the complex features of the local regions. In response to this problem, plenty of subsequent methods have tried introducing context for improvement, with decent results. Yang et al. [32] directly take the global RoI pooling feature as a context and fuse it with region features in LSTM, which provides a promising aid for dense captioning. While Li et al. [15] argue that the caption regions have certain relevance to salient objects at both spatial and semantic conceptual levels, they wish to capture the relevance as context. In particular, Li et al. perform feature extraction of possible RoIs and salient objects and obtain more effective context through feature fusion. Similarly, Yin et al. [35] solve the aperture problem by obtaining contextual information from global features and local features of neighboring RoIs. The above research based on contextual learning has proposed novel solutions for the aperture problem, which

greatly inspires us. However, most methods still adopt a region proposal module with single-scale features to locate salient regions. Thus, the proposals tend to be distributed around several salient regions, resulting in the absence of critical information, which is one of the starting points for our work.

Thanks to the great potential of the transformer for large-scale pre-training tasks, some recent works have attempted to apply the transformer to dense captioning. Shao et al. [24] propose an end-to-end transformer-based framework that considers the relationship between objects and regions and the confidence scores of the detected objects to measure the importance of each region. Wu et al. [28] and Long et al. [21], using large-scale visual and language pre-trained models, both successfully unify open-set object detection and dense captioning within a single framework. These works have greatly advanced the research on dense captioning. However, for unifying object detection and dense captioning, these methods fragment the associations between different regions. Recently, Shao et al. [23] use potentially important textual context and dynamically diversifies the vocabulary bank during the captioning process, thus making the generated captions more diverse, which greatly motivated us. Consequently, we expect to capture and exploit the correlation between regions to assist in the visual understanding of models.

III. METHODOLOGY

As illustrated in Figure 2, we propose a common and distinct knowledge mining network with content interaction for the dense captioning task, CDKM. First, we extract global features from the input images using a pre-trained ViT and generate multi-scale features using a feature pyramid. Next, we adopt a detection network with a knowledge-mining module to detect the salient regions in images and extract the corresponding region features. Furthermore, we calculate the correlation between different regions in the content interaction module to obtain sufficient contextual information. Finally, we use the regional contextual information to guide the caption generator to describe all the salient regions.

A. Multi-scale Feature Extraction

The idea of adopting multi-scale visual features for inference can greatly improve the performance of current object detection methods, which can also apply to dense captioning methods. Inspired by some works [16], [28], we use ViT [4] as the backbone to extract global features of images and then a feature pyramid to acquire multiple visual features at different scales.

In particular, given an image, we divide it into a number of image patches of the same size and convert each patch into a feature vector of fixed length. After that, we employ a pre-trained ViT to learn rich visual features from the input image. Similar to [28], we introduce the window attention mechanism into our backbone to reduce the computational effort and improve efficiency. The window attention mechanism divides the visual features into multiple identical windows and computes the self-attention only within each window. After calculation, each window is remerged and restored to

the original feature map dimensions. The window attention mechanism is incorporated within the several layers of the backbone, while the remaining layers still compute the full self-attention on the feature map. We process the output features of ViT using a feature pyramid [17] to obtain multi-scale features. In detail, the feature pyramid consists of multiple parallel sets of convolution or deconvolution networks, thus mapping the feature map to five different scales, which we denote as $F = \{f_1, f_2, f_3, f_4, f_5\}$. Further details are given in Algorithm 1.

Algorithm 1 Multi-scale Feature Extraction

Input: An Image \mathcal{I}

Output: Multi-scale feature $F = \{f_1, f_2, f_3, f_4, f_5\}$

- 1: If \mathcal{I} is oversized, crop \mathcal{I} to ensure that its sides are no larger than 1024
 - 2: $V_0 \leftarrow$ Implemente patch embedding for \mathcal{I}
 - 3: **for** i in range(1,N+1) **do**
 - 4: $Layer_i \leftarrow$ A ViT layer with full attention if i is divisible by 3 else a ViT layer with window attention
 - 5: $V_i \leftarrow$ Utilize $Layer_i$ to process V_{i-1}
 - 6: **end for**
 - 7: $f_1 \leftarrow$ Expand V_N to twice its size using the Deconvolution and extract features using the Convolution, the scale is $\frac{1}{8}$.
 - 8: $f_2 \leftarrow$ Extract features from V_N using the Convolution, the scale is $\frac{1}{16}$.
 - 9: $f_3 \leftarrow$ Shrink V_N to one-half of its original size using Max Pooling and extract features using Convolution, the scale is $\frac{1}{32}$.
 - 10: $f_4 \leftarrow$ Downsample f_3 to shrink it to one-half of its original size, the scale is $\frac{1}{64}$.
 - 11: $f_5 \leftarrow$ Downsample f_4 to shrink it to one-half of its original size, the scale is $\frac{1}{128}$.
 - 12: $F \leftarrow$ Pack f_1, f_2, f_3, f_4, f_5
 - 13: **return** F
-

B. Common and Distinct Knowledge Mining Network

Depending on the requirements of the dense captioning task, the prominent targets in images are frequently the objects needed to be described. Consequently, existing dense captioning methods [11], [15], [21], [24], [28], [35] tend to locate the salient objects and segment the visual representations of the corresponding areas. However, these methods rely on only a single-scale feature or fail to fully utilize multi-scale features, leading to imperfect performance.

To address the above problem, we consider adequate learning of multi-scale features a key to assisting models in better handling complex scenes containing dense targets. Intuitively, there is both commonality and distinctiveness between visual features at different scales. On the one hand, these visual features all originate from the same image. Thus their internal semantics are quite similar, which is the commonality of multi-scale features. On the other hand, visual features at different scales reflect various perceptual fields, implying diverse semantic levels, which are the distinctiveness of multi-scale

features. We hope to summarize two kinds of knowledge from the commonality and distinctiveness of multi-scale features to assist the model in identifying salient targets. Specifically, we intend to summarize common knowledge that can guide the network to accurately perceive the semantics of the same target when processing visual features at different scales. In addition, we also try to summarize distinct knowledge from each scale of features separately, which guides the network to make fuller use of features based on the characteristics of the current scale.

Based on the above motivation, we design an effective knowledge-mining network, which consists of a shared weight convolution network and multiple independent detection heads. The shared weight convolution network mines the common knowledge of multi-scale features, while the independent detection heads learn the distinct properties of a certain scale feature. To mine deeper common and distinct knowledge, we introduce an implicit knowledge learning mechanism to dig deeper into feature knowledge by utilizing trainable vectors. In the following, we introduce the operation of our module in detail.

The shared weight convolution network needs to process five kinds of features simultaneously to summarize common knowledge. This approach allows the network to ignore the differences among the features and focus on the commonalities of multi-scale features. Taking the first element f_1 in the multi-scale features F as an example, the shared weight convolution network first performs the learning of common knowledge of multi-scale features from f_1 . Specifically, the shared weight convolution network consists of a stack of multiple convolution layers, and the operation process of each layer can be expressed as:

$$f_1^1 = \text{ActF}(\text{Norm}(\text{Conv}(f_1))),$$

where f_1^1 is the output of a single convolution layer whose shape remains the same as f_1 , ActF and Norm are the activation functions and the normalization method. In detail, we use ReLU [7] and GroupNorm [29] as the activation function and the normalization method, respectively.

Through all the convolution layers, we gain the refine feature, which denotes f_1^2 . Further, we introduce an implicit knowledge learning mechanism to simulate latent common knowledge, which is calculated as:

$$f_1^s = (f_1^2 + \mathbb{C}^A) \bullet \mathbb{C}^B,$$

where both \mathbb{C}^A and \mathbb{C}^B are trainable vectors with the same length as f_1 , shared by all the multi-scale features, representing the latent common knowledge. For the input multi-scale features $F = \{f_1, f_2, f_3, f_4, f_5\}$, we set the output of the shared weight convolution network as $F_{share} = \{f_1^s, f_2^s, f_3^s, f_4^s, f_5^s\}$.

For each item in F_{share} , we model the distinct knowledge of various scales with five individual detection heads containing an implicit knowledge learning mechanism. In each detection head, a convolution network is applied to further refine the information about the salient targets. Each detection head only processes single-scale features. Because they are only exposed to features on a single scale, the heads will gradually adapt to

the characteristics of a certain scale. Taking the first detection head as an example, the detailed process can be written as:

$$f_1^d = \text{Head}(f_1^s),$$

$$\text{Head}(\cdot) = \sum_{k=1}^n L_k(\cdot) = \sum_{k=1}^n \text{ActF}(\text{Norm}(\text{Conv}(\cdot))),$$

where f_1^d is the output of the first detection head, n denotes the number of convolution layer stacking in a detection head, and $L_k(\cdot)$ denotes the k -th convolution layer in the detection head. Furthermore, we simulate the distinct knowledge using an implicit knowledge learning mechanism, denoted as:

$$f_1^F = (f_1^d + \mathbb{D}_1^A) \bullet \mathbb{D}_1^B,$$

where both \mathbb{D}_1^A and \mathbb{D}_1^B are trainable vectors for features of the first scale, and f_1^F is the output of the first detection head. By mining and modeling the common and distinct knowledge in multi-scale features, the model is able to utilize multi-scale features in a more efficient manner, and thus locate dense targets more accurately in complex scenes.

The above network will provide proposals for the salient regions. Similar to the method of Zhou et al. [42], we refine these proposals to predict more precise spatial locations for each salient region. This operation is performed by three identical prediction heads. Each prediction head slices the region features based on the proposals by ROI alignment. Subsequently, these region features will be flattened and finally mapped into a 4-dimensional spatial location. Since we assign the tasks of visual understanding and caption generation to the later parts, the network simply produces object proposals without judging the category.

We aggregate and retain all the proposals with high confidence and without overlapping. Besides, we segment the corresponding region information R from the multi-scale features F based on the remaining proposals. Specifically, we match the proposals with different feature maps based on the area of the bounding boxes (i.e., the smaller bounding box and the larger size of the feature map are matched). By means of ROI alignment, we obtain a higher quality set of region features R , each of which has the same dimensions. In addition, to optimize our knowledge-mining network, we adopt binary heatmap focal loss [43] and CIOU loss [40] as the loss functions to improve the model's ability to detect complex targets.

C. Content Interaction Module

Existing transformer-based dense captioning methods tend to feed the region features R directly into the caption generator, thus describing the contents of different regions [21], [28]. This approach means that the caption generator only perceives the local content of the image when describing the salient regions, which definitely severs the potential association between the local areas and the whole scene. Although the content of salient regions is indeed what the dense captioning task focuses on, numerous studies [19], [32] have proved that using global information as a context can facilitate better visual understanding. However, since the relationships between different local regions and the whole scene are complex, it

would be difficult to use a generic context to guide an accurate description when focusing on different regions. To address this problem, we propose a content interaction module that utilizes multiple region features to jointly represent the overall semantics of an image. By calculating the correlation of a salient region with other areas, we can design an exclusive context for this region. The region feature is fed into the caption generator together with its corresponding context, which leads to a more precise description. The details are as follows.

The computational process is shown in Figure 2-(b). Suppose the detecting network proposes m salient regions from the input image and the corresponding region feature set $R = \{r_1, r_2, \dots, r_m\}$. Since each element of R has a large feature dimension, we need to extract the core information of these features to reduce the computation and training difficulty. We perform downsampling for each element in R , thus extracting the core features representing each region from R , which contains considerable redundant information. Taking the first region in R as an example, the operation can be denoted as

$$r_1^c = Conv(r_1),$$

where r_1^c denotes the core feature of the first region conveying critical image information.

Subsequently, we aggregate the core features of different regions, denoted as R_C , and compute the self-attention on R_C . First, we use a linear layer to map R_C to obtain a key vector R_{key} containing the information for all the proposed regions. Then, we calculate the correlations between different regions as the basis for computing contextual information. Taking the correlation computation of the first region as an example, we multiply the core feature of this region, r_1^c , with the transpose of R_{key} . We apply *softmax* to multiplication results as a way to obtain the quantified internal correlations between the first region and the other proposed regions. The above calculation can be written as:

$$corl_1 = softmax\left(\frac{r_1^c}{\sqrt{d_1}} \bullet R_{key}^T\right),$$

where $corl_1$ denotes the attention score of the first region to all the proposed regions, and d_1 denotes as the length of r_1^c .

Based on $corl_1$, we adjust the weights of various regional influences and model the global information using all weighted region features. In order to limit the excessive influence introduced by other regions, we introduce the core features of the first region in the calculation of the global information. The above process is denoted as:

$$Global_1 = fc(corl_1 \bullet fc(R_C) + r_1^c),$$

where fc denotes a linear layer, $Global_1$ means the exclusive global information for the first region.

Finally, we augment the global information with a multi-layer perceptron and generate the context of the first region in the form of a visual token, which can be formulated as:

$$ConToken_1 = MLP(Global_1) + Global_1,$$

where MLP denotes the multi-layer perceptron, and $ConToken_1$ denotes the correlation context for the first

region. The context will serve as a visual token that is concatenated with the corresponding region features to form a new visual combination for the region caption generation. In detail, the process is denoted as:

$$r_1^f = Concat(ConToken_1, r_1),$$

where r_1^f denotes the final visual feature of the first region, and we define the final region features set $R_F = \{r_1^f, r_2^f, \dots, r_m^f\}$. The deliberate context tokens can provide more relevant global information to the local areas during the description process, allowing the model to refine its region understanding based on the global information, thus improving the quality of the captions.

The BERT-based caption generator, which consists of a 6-layer network with a hidden dimension of 768, takes R_F as input and describes the proposed regions in an autoregressive manner. We follow Wu et al. [28] and reorganize each region feature into a new batch, which enables the caption generator to handle an indeterminate number of regions in parallel. We mitigate the aperture problem at this stage through our correlation context. In the training stage, each proposed region is matched with its closest ground truth box, and the region features in R_F are concatenated with the text features from their corresponding ground truth captions. We optimize the model by using a cross-entropy function that calculates the loss between the predicted captions and the ground truth.

IV. EXPERIMENTS

A. Datasets

We investigate a wide range of datasets involved in current research on dense captioning and conduct sufficient experiments on these datasets. In particular, these datasets include Visual Genome V1.0 and V1.2 [12], and the VG-COCO dataset [15].

Visual Genome (VG). The VG dataset is dedicated to modeling potential interactions and relationships between various objects in an image and contains dense annotations, attributes, and relationships among the objects. We have selected the dense object annotations within the VG dataset for training and testing and divided the dataset following the approach of [11]. In detail, a total of 77,398 images are used for training and 5,000 images for validation and testing. In addition, the VG dataset is currently available in three versions: V1.0, V1.2, and V1.4. V1.2 and V1.4 have the same dense object annotations and are richer compared to V1.0; therefore, we conduct experiments on V1.0 and V1.2, respectively.

VG-COCO. VG-COCO is the intersection of VG V1.2 and MS COCO [18] proposed by Li et al. The training split, validation split, and test split of the dataset have 38,080, 2,489, and 2,476 images, respectively. Li et al. point out that the VG dataset has denser bounding boxes than other object detection benchmarks. For instance, the average number of objects per image within the VG V1.2 training set is 35.4, while the MS COCO is only 7.1. To compensate for the problem of overly dense bounding boxes within the VG dataset and to provide proper bounding boxes and dense object annotations for the dense captioning task.

TABLE I: Comparison results of our proposed method with existing methods of dense captioning. We classify the existing methods into two categories, i.e., CNN-and-RNN-based and transformer-based. The experiments are conducted on three datasets, and the evaluation metric adopted is the dense captioning mAP (%).

Type	Method	VG V1.0	VG V1.2	VG-COCO
CNN+RNN	FCLN (2016) [11]	5.39	5.16	4.23
	JIVC (2017) [32]	9.31	9.96	7.85
	ImgG (2019) [15]	9.25	9.68	7.81
	COCD (2019) [15]	9.36	9.75	7.92
	COCG (2019) [15]	9.82	10.39	8.90
	CAG-Net (2019) [35]	-	10.51	-
transformer	TDC+ROCSU (2022) [24]	11.49	11.90	11.58
	CapDet (2023) [21]	-	15.44	13.98
	GRiT (2022) [28]	15.52	-	-
	ETDC+TCM+DVFH (2023) [23]	13.24	12.60	14.30
Our Method		15.82	16.15	16.00

B. Implementation Settings

In the knowledge mining network, we filter the proposals to retain non-overlapping results using NMS with a threshold of 0.9. Up to 2000 results with the highest confidence will be retained in the training phase and up to 256 in the testing phase. We adopt the AdamW optimizer to train the CDKM on RTX A40, setting the batch size to 4, the initial learning rate to 0.00008, and the number of iterations to 180,000. We test on multiple datasets. For the VG dataset, we train CDKM following the dataset partitioning approach of Johnson et al [11]. Specifically, 77,398 images are used for training and 5,000 for testing. For the VG-COCO dataset, we follow the dataset partitioning approach of Li et al [15]. 38,080 images are used for training and 2,476 images for testing.

C. Evaluation Metric

For a fair comparison, we adopt the same evaluation metric, mean Average Precision (mAP), as the vast majority of current methods. This evaluation metric, which is specific to the dense captioning task, considers both the ability to locate salient targets and the capability to describe the content of a region. To distinguish it from the mAP used for object detection, we denote it as dense captioning mAP. Specifically, we first preprocess the ground truth bounding boxes, combine overlapping ground truth bounding boxes (i.e., IoU less than 0.7) into one larger area, and combine the corresponding captions of these boxes. We reorder the prediction bounding boxes for each image in descending order of confidence and calculate their IoU with all the ground truth bounding boxes. We find the closest GT bounding boxes for each prediction. If more than one prediction matches the same GT bounding boxes, only the prediction with the highest confidence is retained. After the matching is completed, we compute the Meteor scores [3] for each generated caption with the matched ground truth to serve as a criterion for evaluating the language similarity. Following the threshold settings of Johnson et al. [11], we use different IoU thresholds (0.3, 0.4, 0.5, 0.6, 0.7) and different Meteor score thresholds (0, 0.05, 0.10, 0.15, 0.20, 0.25) to measure the average precision across all pairwise settings. Finally, we

take the means of all the average precisions to measure the effectiveness of the methods.

D. Comparison with the State-of-the-Art Methods

In this section, we compare our method with previous State-of-the-Art methods on three datasets, VG V1.0, VG V1.2, and VG-COCO, which are reported in Table I. Specifically, the compared benchmarks include FCLN [11], JIVC [32], ImgG [15], COCD [15], COCG [15], CAG-Net [35], TDC+ROCSU [24], GRiT [28], CapDet [21] and ETDC+TCM+DVFH [23]. As we can see, these methods are classified into two categories. Earlier methods for dense captioning generally consist of CNNs and RNNs, which have certain shortcomings in their performance on various datasets. In recent years, transformer-based methods have gradually become the research focus of dense captioning, with superior performance. This is because these methods make full use of pre-trained vision and language models, resulting in more efficient image representations and accomplishing the given task through more powerful linguistic expressions.

As illustrated in Table I, our method outperforms existing methods on all three datasets. First, our method achieves mAP scores of 15.82% and 16.15% in the VG V1.0 and V1.2 datasets with more complex scenarios, respectively. This implies that our method can better cope with the dense captioning task in scenarios with a large number of targets. In addition, it is worth pointing out that our method, GRiT and CapDet achieve superior results compared to TDC, although they are all based on the transformer architecture. This is due to the fact that all these methods employ powerful pre-trained vision models as their backbone, resulting in richer visual representations. However, since GRiT and CapDet try to implement open-set object detection with dense captioning under a unified framework, they have to spare loading more visual understanding mechanisms in the decoders. In contrast, our method focuses on dense captioning, introducing additional valid context when describing salient regions, which reinforces the model's image understanding. Furthermore, our method significantly improves the mAP score from 13.98% to 16.00% on the VG-COCO dataset, which indicates that our method can still accurately capture salient objects when there are relatively few targets. In total, the excellent performance is attributed to our common and distinct knowledge-mining network, which fully learns and utilizes multi-scale visual features, resulting in higher sensitivity to salient objects. For our common and distinct knowledge-mining network and content interaction module, we conduct more in-depth experiments and analyze their roles below.

E. Ablation Study

1) *The Effect of the Common and Distinct Knowledge Mining Network:* In this section, we conduct an ablation study of the common and distinct knowledge-mining network so as to demonstrate the contribution of our network. Specifically, we divide the network into three parts, i.e., the shared weight convolution network, the multiple detection heads, and the implicit knowledge learning mechanism, and examine the actual



Fig. 3: The qualitative comparisons between our proposed method and GReT on the VG v1.2 test split. For a clearer comparison, we use blue, red, and green to represent the bounding boxes from ground truth, GReT, and our method, respectively. In addition, we report the IOU and METEOR scores between the predicted results and the ground truth.

TABLE II: The ablation study on the common and distinct knowledge-mining network. We divide the network into three parts and delve into the role of each component. Here, we use both dense captioning mAP and detection mAP to measure the performance of different schemes, denoted as Dcap mAP and Det mAP, respectively.

No	Shared weight layer	Multi detection heads	Implicit knowledge	VG V1.2	
				Dcap mAP (%)	Det mAP (%)
1	✓	✗	✗	14.08	26.39
2	✗	✓	✗	14.27	27.06
3	✓	✓	✗	15.77	28.20
4	✓	✗	✓	14.15	26.70
5	✗	✓	✓	14.49	27.06
6	✓	✓	✓	16.15	28.64

effects of each component on salient target detection (detection mAP) and dense captioning (dense captioning mAP). The experiments are conducted on the VG V1.2 dataset, and the results are reported in Table II.

First, compared with No. 4, No. 5, and No. 6, it is clear that our method is much better than the other two structures.

The reason is that these two structures fail to acquire both common and distinct knowledge of multi-scale features. Our knowledge-mining network indeed captures the commonality and distinction of the multi-scale features, which achieves adequate knowledge learning. Second, when compared to other methods, No. 6 gains a higher detection mAP score. This demonstrates that the proposed module has a more powerful target localization capability, which in turn improves dense captioning. Finally, as illustrated in Table II, we implement the implicit knowledge learning mechanism in Nos. 4, 5, and 6, with better detection mAP and dense captioning mAP scores. This is due to the implicit knowledge learning mechanism, which further enhances the salient region proposal module's ability to utilize multi-scale features, thereby facilitating the performance of dense captioning.

2) *The Effect of the Content Interaction Module:* We delve into the role of the content interaction module. Specifically, we compare the effects of four kinds of contextual information on the common captioning framework (i.e., only using a shared weight convolution network for proposal) and our proposed framework, respectively, so as to demonstrate the

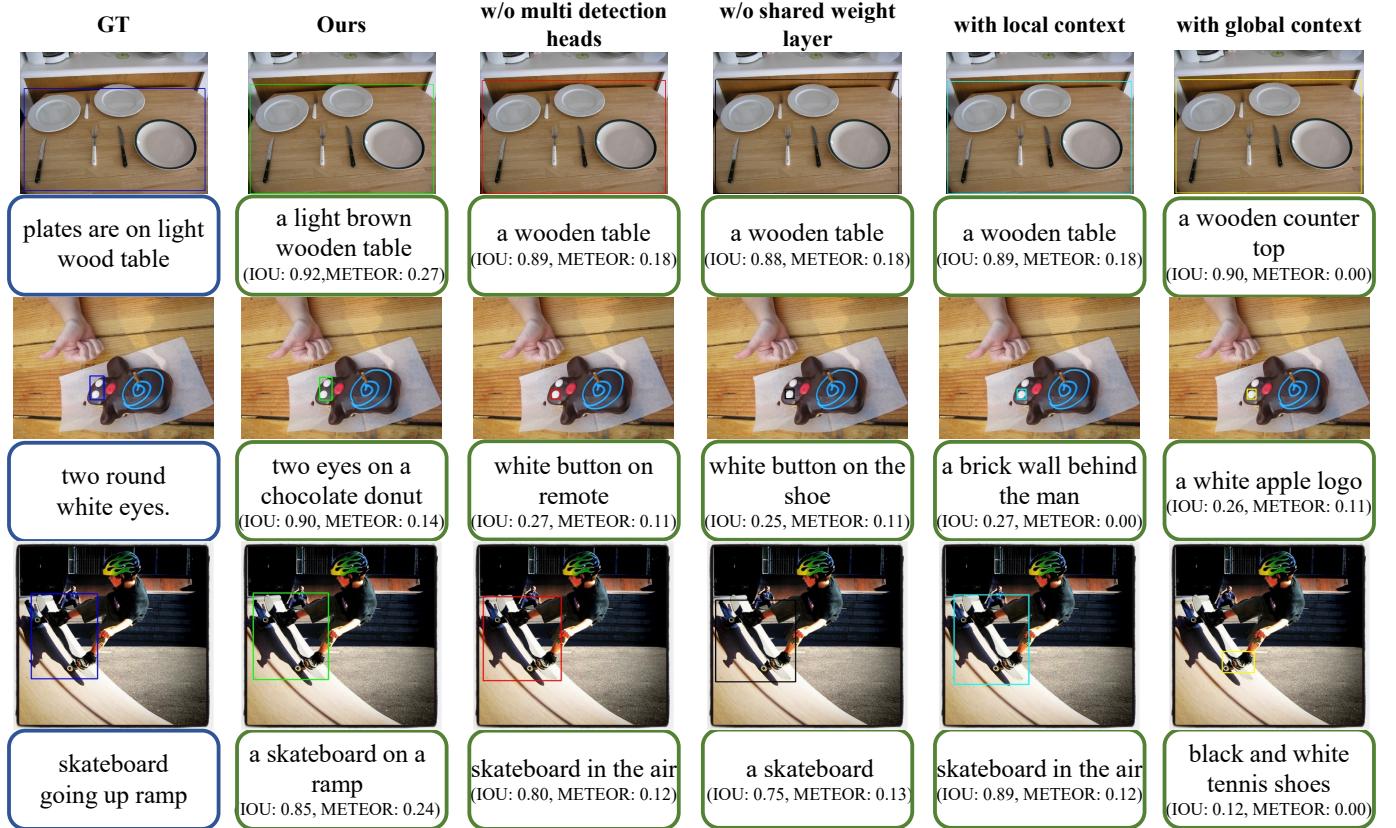


Fig. 4: The qualitative comparisons of ablation results on the VG v1.2 test split. We compare our method with several ablation schemes. These ablation schemes include removing multiple detection heads, removing shared weight convolution network, adopting local context and adopting global context, respectively.

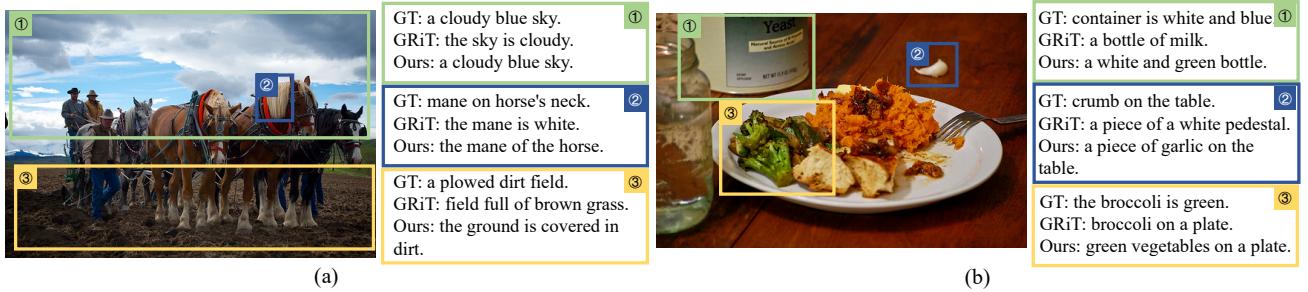


Fig. 5: Some of the complete visualization results on the VG v1.2 test split. For each image, we select three different regions and display captions from ground truth, GRiT, and our methods.

effectiveness of our proposed module. Further, the four kinds of contextual information include the blank context (a zeros-tensor), the global context (the feature maps from ViT), the local context (the core features R_C), and our correlation context $ConToken$, respectively. The experiment results are presented in Table III.

First, the comparison between No. 1 and No. 5 indicates that the ability of dense captioning is relatively poor when no contextual information is used, although the region proposal is not significantly affected. This suggests that contextual information facilitates the visual understanding of localized regions, contributing to better captioning. Furthermore, we find that

our correlation context has superior results to other contextual information. No. 4 and No. 8 show an increase in the model's dense captioning mAP of 1.09% and 2.47%, respectively. This is because our correlation context has provided effective surrounding information to the caption generator. In contrast, No. 2 and No. 6 show that the methods with global context instead become less effective. The reason is that transformer-based methods have a strong regional understanding and traditional global contexts only introduce more distracting information. This also justifies the need to apply exclusive contexts for different regions. In addition, it is noteworthy that No. 3 has achieved similar results to No. 4. We argue that this is due

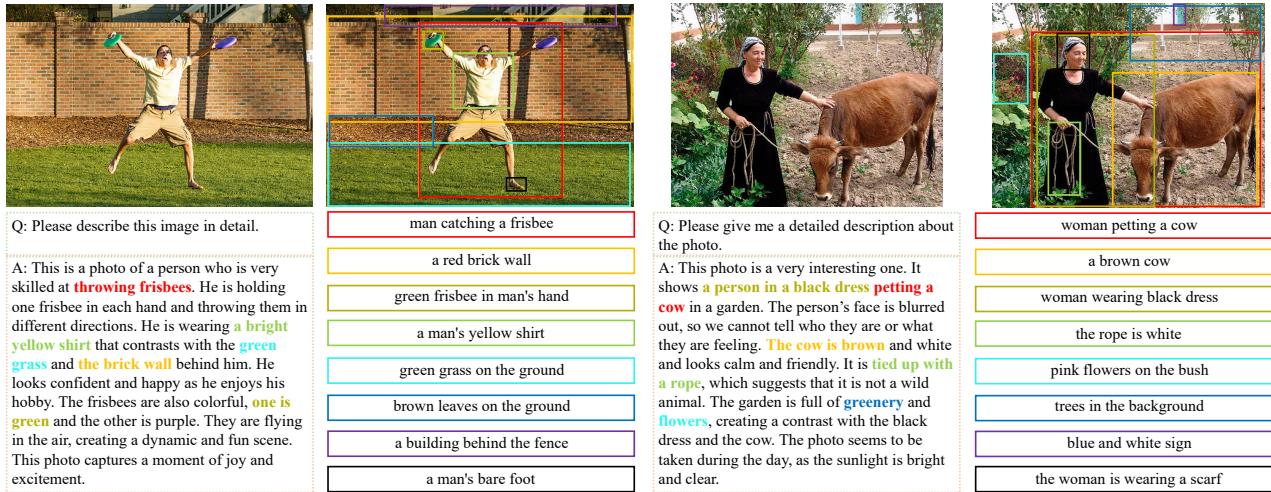


Fig. 6: Some results of comparison with GPT-4. Our proposed dense captioning method gives a better detailed description of the images than GPT-4. More results are detailed in the Appendix.

TABLE III: The ablation study on the content interaction module. Similar to Table II, we take dense captioning mAP (Dcap mAP) and detection mAP (Det mAP) as evaluation metrics. In the ablation study, we use four different kinds of contextual information, i.e., the blank context, the global context, the local context, and our correlation context, denoted as Blank, Global, Region, and Correlation, respectively.

No	Multi detection heads	Addition info	VG V1.2	
			Dcap mAP (%)	Det mAP (%)
1	x	Blank	13.06	26.89
2	x	Global	13.74	26.19
3	x	Region	14.05	26.51
4	x	Correlation	14.15	26.70
5	✓	Blank	13.68	28.17
6	✓	Global	14.62	27.23
7	✓	Region	14.98	27.98
8	✓	Correlation	16.15	28.64

to common captioning frameworks' relatively poor detection performance, leading to biases in the proposed salient regions. This, in turn, results in the context module misleading the visual understanding of the salient regions.

F. Qualitative Results and Analysis

In this section, to demonstrate the superiority of the method more directly, we visualize some results generated by our method to qualitatively evaluate it.

1) *Qualitative Comparisons with Baseline*: Figure 3 displays the different performances of our method and the SOTA method, GRiT, on several samples. We are able to draw two conclusions from the comparisons. First, it is obvious that our method has better performance in both target detection and content description. Taking Figure 3-(a) as an example, for the motorcycle on the left side, our method recognizes the target more perfectly, while the closest recognition result of GRiT still has a large deviation. Secondly, our method utilizes information surrounding the target to enhance the visual understanding, thus providing a more detailed and comprehensive caption. For example, Figure 3-(b) shows that

when both methods describe the same region, our method attends to the sharp distribution of the different color blocks, thus tending to offer a more detailed caption with a higher METEOR score. Furthermore, in Figure 3-(c), the results of our method are closer to the ground truth in terms of both region location and description content. Moreover, in Figure 3-(d, e, f), our model combines the content of the associated regions (i.e., the fences) when describing the target region, leading to a more comprehensive caption. In summary, the examples above have demonstrated the effectiveness of our method.

2) *Qualitative Comparisons of Ablation Results*: In order to more intuitively represent the effect of our proposed common and distinct knowledge-mining network and content interaction module, we compare the visualization results of several ablation experiment schemes. Specifically, these ablation experiment methods include removing multiple detection heads, removing shared weight convolution networks, and adopting local context, and global context. As shown in the Figure 4, all methods describe the same regions. The results in the first row indicate that our method generates captions with richer details. This is due to the coordinated effect of the knowledge-mining network and the content interaction module that enhances the model's ability to recognize the images. Furthermore, in the second row, we notice that the captions in columns 2, 3, and 4 all express the meaning of white eyes placed on other objects to some extent. Although this is somewhat different from the ground truth, we consider the performance worthy of recognition, as it provides users with more comprehensive information. The reason is that the content interaction module provides information about relevant objects when perceiving tiny objects, thus enhancing the visual understanding of the local area. In contrast, the results in columns 5 and 6 both deviate substantially from the ground truth without the correlation information. Lastly, in the third row, our method again provides a more comprehensive caption, which is more similar to the ground truth obviously.

This is attributed to the knowledge-mining network accurately capturing various objects (e.g., skateboard, ramp), which lays the foundation for the model's robust visual understanding.

3) *Other Visualization Comparisons*: Finally, there are some other visualization comparisons. To begin with, we provide some complete visualization results, in which we describe the three different regions of each image, as a way to illustrate the difference between our method and the baseline in practice. Figure 5-(a) demonstrates that our method achieves a more accurate description for multiple regions of the scene. Taking area 3 as an example, our method successfully cognizes the concept of the dirt field, whereas GRiT incorrectly recognizes the dirt field as brown grass. In area 2, although both methods are able to cognize the concept of mane, our method is able to provide a more comprehensive caption, i.e., to illuminate the association of mane with the horse, thanks to the fact that it effectively combines information from the surrounding area. These same conclusions mentioned above are also reflected in Figure 5-(b). However, there remain a number of problems with our method that need to be addressed. For ambiguous objects in complex scenes, our method still struggles to accurately recognize them. Taking area 2 in Figure 5-(b) as an example, misidentification occurs for both powerful methods for the piece of crumb. Even so, our method acquires surrounding information related to the food, thus refining the visual understanding and providing a relatively relevant answer. This further supports the necessity of introducing relevant regional information as context in dense captioning. In summary, our method can provide users with more detailed and comprehensive captions and therefore has higher application value and effectiveness.

Furthermore, we likewise attempt to demonstrate the value of our method by contrasting it with the latest language models. Specifically, we guide GPT-4 to provide a detailed description of a given image and briefly compare the richness of the information provided by the two approaches, as shown in Figure 6. It can be seen that GPT-4 works well to provide the user with a vivid description based on the overall content of the image and to convey the emotion that lies behind the image. Despite this stunning performance, GPT-4 inevitably omits some sub-important information when describing the picture. Taking the left side of Figure 6 as an example, our method provides the caption of the primary targets while still capturing other sub-important targets such as dead leaves, bare feet, buildings, etc., which are missed by GPT-4. This illustrates the fact that current language models still fail to replace our method of dense captioning and demonstrates the research value of the dense captioning task. It is confident that the research on dense captioning will further contribute to the advancement of multimodal models in the future.

V. CONCLUSION

In this work, we propose a common and distinct knowledge-mining network with content interaction for the dense captioning that achieves the state-of-the-art on multiple datasets. Specifically, in order to better utilize multi-scale features for salient target detection, we propose a novel knowledge-mining

network for mining the common and distinct knowledge of multi-scale features. In addition, we propose a novel contextual representation based on regional correlation to generate exclusive contextual information for different salient regions, which improves the model's visual understanding of local regions. Through sufficient quantitative and qualitative experiments, we validate the effectiveness of these modules. However, there is still much room for improvement in our approach. For example, our utilization of regional correlations is still not sufficient and may interfere with the model's understanding and captioning of regions in some cases. In future work, we will continue to investigate more effective methods for calculating regional correlations and better assisting dense captioning by providing more effective contextual information.

REFERENCES

- [1] Nayyer Aafaq, Ajmal Mian, Naveed Akhtar, Wei Liu, and Mubarak Shah. Dense video captioning with early linguistic information fusion. *IEEE Transactions on Multimedia*, 25:2309–2322, 2023.
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei 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*, pages 6077–6086, 2018.
- [3] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [5] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, pages 266–282. Springer, 2022.
- [6] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.
- [7] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.
- [8] Wided Hammadi, Metzli Ramirez-Martinez, Philippe Brunet, Sidi-Mohamed Senouci, and Mohamed Ayoub Messous. Deep learning-based real-time object detection in inland navigation. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2019.
- [9] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- [10] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019.
- [11] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.
- [12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [13] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
- [14] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 21(8):2117–2130, 2019.

- [15] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8650–8657, 2019.
- [16] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.
- [17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [19] Chencheng Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *CVPR*, June 2020.
- [20] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128:261–318, 2020.
- [21] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. Capdet: Unifying dense captioning and open-world detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15233–15243, 2023.
- [22] Aldosary Saad and Abdallah A Mohamed. An integrated human computer interaction scheme for object detection using deep learning. *Computers & Electrical Engineering*, 96:107475, 2021.
- [23] Zhuang Shao, Jungong Han, Kurt Debattista, and Yanwei Pang. Textual context-aware dense captioning with diverse words. *IEEE Transactions on Multimedia*, pages 1–15, 2023.
- [24] Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [25] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. Image captioning: a comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pages 325–328. IEEE, 2020.
- [26] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.
- [27] Qi Wang, Hongyu Deng, Xue Wu, Zhenguo Yang, Yun Liu, Yazhou Wang, and Gefei Hao. Lcm-captioner: A lightweight text-based image captioning method with collaborative mechanism between vision and text. *Neural Networks*, 162:318–329, 2023.
- [28] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [29] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [30] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020.
- [31] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):43–51, 2022.
- [32] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017.
- [33] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, 21(4):1047–1061, 2019.
- [34] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.
- [35] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6241–6250, 2019.
- [36] Litao Yu, Jian Zhang, and Qiang Wu. Dual attention on pyramid feature maps for image captioning. *IEEE Transactions on Multimedia*, 24:1775–1786, 2022.
- [37] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.
- [38] Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Luping Zhou. Dense video captioning using graph-based sentence summarization. *IEEE Transactions on Multimedia*, 23:1799–1810, 2021.
- [39] Zongjian Zhang, Qiang Wu, Yang Wang, and Fang Chen. High-quality image captioning with fine-grained and semantic-guided visual attention. *IEEE Transactions on Multimedia*, 21(7):1681–1693, 2019.
- [40] Zhaozhui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.
- [41] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.
- [42] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021.
- [43] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.



Hongyu Deng is currently pursuing his master's degree in State Key Laboratory of Public Big Data, Guizhou University, China. His research interests include artificial intelligence, computer vision, multimodal fusion and image captioning.



Yushan Xie is currently pursuing her master's degree at the School of Computer Science and Technology, Guizhou University, China. Her research interests include artificial intelligence, computer vision, multimodality and image captioning.



Qi Wang (Member, IEEE) is currently a Special Term Professor in State Key Laboratory of Public Big Data, Guizhou University. He received B.E. degree from Northwest nationality University, Gansu, China, in 2015, and Ph.D. degree in Computer Application Engineering at School of Computer Science and Technology, Guangdong University of Technology, Guangdong, China, in 2020. He also received a Ph.D. degree in Engineering Technology at Faculty of Engineering and Technology, Hasselt University, Hasselt, Belgium, in 2021. His current research interests include computer vision, AI security, agricultural vision and text and vision scope. He has authored or co-authored over 20 papers in prestigious conferences and journals in computer vision and multimedia.



Jianjun Wang is currently pursuing his master's degree in State Key Laboratory of Public Big Data, Guizhou University, China. His research interests include artificial intelligence, computer vision, fine-grained image classification.



Weijian Ruan is currently the President of Innovation R&D Development in Smart City Research Institute of CETC, and had won four great awards like Guangdong Provincial Science and Technology Progress Award. He received his B.E. and Ph.D degrees from Wuhan University, China, in 2014 and 2020, respectively. He had ever been a long-term researcher in National Institute of Informatics, Tokyo, Japan, JD AI Research, and Pengcheng Laboratory from 2018 to 2020. His research interests include computer vision and multimedia analysis, where he

has published more than 30 papers, including IEEE TCYB/TMM/TCSVT, ICCV, ACM MM, AAAI, etc.



Wu Liu (Senior Member, IEEE) received the B.E. degree from Shandong University, Shandong, China, in 2009, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China in 2015. He is currently a Special Professor at the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. Before that, he was a Senior Researcher in JD Explore Academy, China from 2018 to 2024, and a Lecturer at the

Beijing University of Posts and Telecommunications from 2015 to 2018. His current research interests include multimedia analytics, search, and generation. He received the IEEE Trans. on Multimedia 2019 Prize Paper Award, IEEE Multimedia 2018 Best Paper Award, and IEEE ICME 2016 Best Student Paper Award. He was elected as the IET Fellow in 2024, 2022 MIT Technology Review 35 Innovators Under 35 Asia Pacific, and received the 2021 ACM China Rising Star Award, the Tianjin Science and Technology Progress Special Award, the Chinese Academy of Sciences Outstanding Ph.D. Thesis Award, etc. He is or has been an Associate Editor of IEEE Trans. on Multimedia from 2022 to 2026, and the Program Co-Chairs of IEEE ICME 2022 and ACM MM Asia 2021.



Yong-Jin Liu received the B.Eng. degree from Tianjin University, Tianjin, China, in 1998, and the M.Phil. and Ph.D. degrees from The Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. He is currently a Professor with the BNRIst, Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include computational geometry, computer vision, cognitive computation, and pattern analysis.

APPENDIX A SUPPLEMENTARY MATERIALS

In order to compare CDKM and GPT-4 more comprehensively, we have added experiments and carried out in-depth analyses to demonstrate the significance of the CDKM method and even dense captioning research. Specifically, Figure A1 shows that CDKM can attend to some hidden details and reflect the spatial information of the objects, Figure A2 demonstrates that CDKM gives an objective description.



Q: please describe the image

A: The image is a creative capture using a convex mirror typically found in subway stations. The mirror reflects a man taking a photograph of himself and the scene around him. He's wearing a bright orange jacket, standing out in the reflection against the subway station backdrop. The mirror distorts the perspective, bending straight lines and enlarging objects in the center of the reflection. In the reflection, you can also see the interior of a subway car with passengers sitting inside; the car is decorated with a blue stripe on the outside. The mirror is emblazoned with the words "THE TUBE," a nickname often used for the London Underground. The setting looks to be a subway platform, with typical underground station elements such as tiled walls and fluorescent lighting. The overall feel is dynamic and somewhat urban-quirky, given the distortion of the mirror and the photographer's self-portrait attempt.

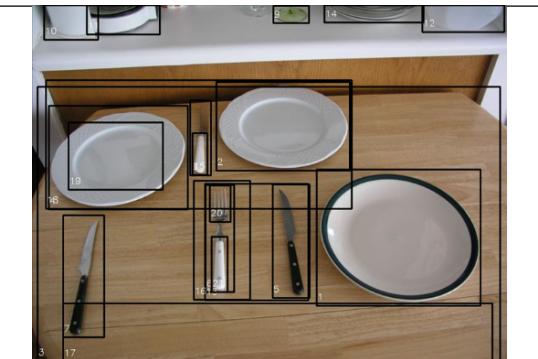


(a)



Q: Can you give me a description for the picture?

A: The image shows a wooden dining table set for two. On the table, there are two white ceramic plates, each accompanied by a knife on the right side, but only one plate has a fork on the left. The plate in the foreground features a simple design with a single green band around the rim, while the other two plates have a more intricate design with embossed patterns along their edges. The cutlery has black handles, except for the fork, which has a white handle that contrasts with the others. Behind the table setting, there's a cabinet or hutch with several shelves that contain various other dining ware items like stacked plates and possibly cups or bowls, but the details are a bit unclear due to the perspective. The table setup appears casual, and the open cabinet suggests an everyday home environment. The scene is likely a depiction of a simple, domestic mealtime setting.



(b)

Fig. A1: Some results of comparison with GPT-4.



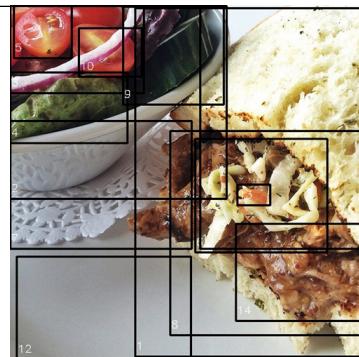
Q: please describe the image

A: This image features a group of cyclists waiting at a traffic intersection. They are dressed in what appears to be vintage or retro attire, suggesting a themed ride or event. The clothing style includes tweed jackets, flat caps, and vests, reminiscent of early to mid-20th century fashion. The bicycles also look to be of various vintage designs, with some having baskets and others equipped with luggage racks. The scene is set against a backdrop of classic and modern architecture, indicating this could be in an urban area with a mix of historical and contemporary buildings. Traffic lights are visible, showing red, instructing the cyclists to stop. The weather seems to be overcast, common in settings like London.



1	man wearing a brown jacket	2	a tan hat on a man's head
3	man riding a bicycle	4	man wearing gray jacket
5	a basket on the front of a bicycle	6	a crowd of people on the street
7	man wearing a blue jacket	8	a large white building
9	a large white building	10	man wearing a blue and white striped shirt
11	a man riding a bicycle	12	a white sign on the back of a bike
13	a tall black street light	14	red line on the road
15	man wearing a brown jacket	16	
17		18	
19		20	
21		22	

(c)



Q: Can you describe this photo?

A: This photo shows a close-up of a sandwich and a side salad, arranged for a meal. The sandwich is cut in half, revealing a hearty filling of what appears to be pulled or shredded meat, topped with slaw or a mix of shredded vegetables. It's on a type of white bread that looks fluffy, with a lightly toasted surface. Next to the sandwich, there's a small bowl containing a salad, which includes colorful ingredients like cherry tomatoes, red onion rings, and leafy greens that could be lettuce or spinach. The bowl seems to be embossed with a decorative pattern, and the salad adds a fresh contrast to the hearty sandwich. The food is presented on a clean, white surface, with a doily under the sandwich adding a touch of elegance to the plating. The image suggests a focus on flavor and texture, combining the savory taste of the meat with the fresh crispness of the salad.

1	a sandwich on a plate	2	a bowl of vegetables
3	sliced red tomatoes	4	a piece of green lettuce
5	a red tomato slice	6	onions on the sandwich
7	white paper napkin under bowl	8	a sandwich on a white plate
9	a green pickle spear	10	a slice of tomato
11	the bread is white	12	the plate is white
13	a piece of carrot	14	meat on the sandwich
15		16	
17		18	
19		20	
21		22	

(d)

Fig. A2: Some results of comparison with GPT-4.