OSIC: A New One-Stage Image Captioner Coined

Bo Wang[1] , **Zhao Zhang**[1*] , **Mingbo Zhao**[2] , **Xiaojie Jin**[3] , **Mingliang Xu**[4] , **Meng Wang**[1]
[1]Hefei University of Technology, Hefei, China
[2]Donghua University, Shanghai, China
[3]Bytedance Research, USA
[4]Zhengzhou University, Zhengzhou, China
{runbor1993, cszzhang}@gmail.com, mzhao4@dhu.edu.cn,
jinxiaojie@bytedance.com, iexumingliang@zzu.edu.cn, eric.mengwang@gmail.com

Abstract

Mainstream image captioning models are usually two-stage captioners, i.e., encoding the region features by a pre-trained detector and then feeding them into a language model to generate the captions. However, such a two-stage procedure will lead to a task-based information gap that decreases the performance, because the region features in the detection task are suboptimal representations and cannot provide all the necessary information for subsequent captions generation. Besides, the region features are usually represented from the last layer of the detectors that lose the local details of images. In this paper, we propose a novel One-Stage Image Captioner (OSIC) with dynamic multi-sight learning, which directly transforms the images into descriptive sentences in one stage for eliminating the information gap. Specifically, to obtain rich features, multi-level features are captured by Swin Transformer, and then fed into a novel dynamic multi-sight embedding module to exploit both the global structure and local texture of input images. To enhance the global modeling capacity of the visual encoder, we propose a new dual-dimensional refining to non-locally model the features interaction. As a result, OSIC can directly obtain rich semantic information to improve the captioner. Extensive comparisons on the benchmark MS-COCO, Flickr8K and Flickr30K datasets verified the superior performance of our method.

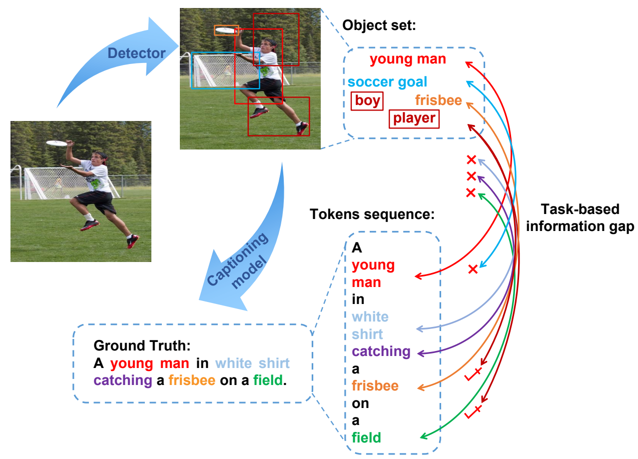

Figure 1: Illustration of the task-based information gap. The language model generates sentences conditioned on the region features of the given image from a fixed detector. The red, orange and blue marked young man, frisbee, and soccer goal are detected by the fixed detector. However, the detection-based region features may not suit for the image captioning task. E.g., the scene "*field*" and object relation "*catching*" are ignored by the detector, without adaption to the descriptive contents of the image. The detected soccer goal is noise for the captioning task. The boy and player marked in dark red are redundantly detected for the sentence. The detected objects are mutually independent, which loses the semantic relationship.

1 Introduction

Image captioning is an important cross-modal task of automatically generating the descriptions of the main contents for given images. Inspired by the procedure for neural machine translation, encoder-decoder architecture is most widely used for image captioning [Cornia *et al.*, 2020], which encodes the given image into the intermediate representation via a vision encoder, followed by a NLP decoder to generate the captions. As a result, the performance of the captioner largely relies on the representations containing information for the NLP decoder. Motivated by this end, the early work aim to compress the image into the fixed-length vector [Vinyals *et al.*, 2015] as visual features. To enrich the compact expressions, grid features [Zhang *et al.*, 2021] are further generated by the CNNs to embed more visual information. More recently, compared with the grid features, two-stage captioners using region features [Anderson *et al.*, 2018] have made great progress by capturing the salient region-level features.

It is noteworthy that the region features obtained by a fixed detector (e.g., Faster R-CNN [Ren *et al.*, 2015]) focus on the detection task, which means it can hardly provide all necessary descriptive information for the subsequent image captioning task [Kuo and Kira, 2022], due to the large information gap between the two tasks. In other words, the vi-

[*]Corresponding author.

sual encoder in the first stage of those captioners is optimized using the detection tags instead of sentences, which causes the information mismatch in feature embedding. We call this mismatch between detection-centric features and captioning-guided features ***task-based information gap***. This gap limits the model to obtain a global optimization, and results in two major issues. Firstly, region features can hardly present all the necessary descriptive information for the target captions. For example, misdetection, insufficient detection (e.g., the scene "*field*" and the object relation "*catching*" are not detected), or redundancies (e.g., "*soccer goal*" and "*player*") are produced, as shown in Figure 1. And more to the point, these region features are represented independently [Hu *et al.*, 2018], without visual semantic connection with each other. But as a text sequence, the captions have clear semantic order assigned for each word [Liu *et al.*, 2017]. Secondly, region features are usually presented by the deepest pooling features [Ren *et al.*, 2015], which may lose the local details. For example, the man's detail "*white shirt*" is missing, and the inadequate descriptions will also decrease the captioning performance.

In this paper, we integrate the intermediate representation and captions generation into a novel one-stage trainable model to obtain a globally optimal solution. The representation misalignment between region features and descriptive semantic features comes from the annotation gap between the detection and captioning, which is essentially the task-based information gap. So, we believe that only the text-annotated optimization based on the one-stage captioner or the collaborative optimization on multi-tasks can close the task-based information gap. Hence, we solve the problem with a one-stage captioner due to its lower annotation cost. In addition, these methods use the embedded features of fixed sight, so they cannot flexibly capture effective visual information of different sizes as well as discover the visual relationships of different distances. To address the above problems, we embed the visual features using the multi-level output of the Swin Transformer (SwinT). We then calculate the dynamic correlation between the features of different sights, so as to grasp the interconnected visual representation. Finally, we consider refining the features on both spatial and channel dimensions to improve the global representation capacity of the encoder for generating richer and more accurate captions. Overall, the main contributions of the paper are summarized as follows:

- We first clearly define the task-based information gap in captioners as the representation mismatch between detection-centric and captioning-guided features. We then propose a novel one-stage image captioner (OSIC) with a dynamic multi-sight learner, which is dedicated to optimizing the captioning framework and visual representations to eliminate the gap for image captioning.
- We propose a dynamic multi-sight embedding to adaptively capture and fuse the global structure in large sight and local texture in small sight. Specifically, it computes the salience coefficient of the embedded features in different sights to embed the multi-sight information dynamically, based on the long and short distance dependences of the Swin Transformer.
- In order to improve the global-interactive ability of the SwinT, we propose a dual-dimensional refining to non-locally enable the features interaction in spatial and channel dimensions, so that the global representation ability of the encoder can be fully enhanced.

2 Related Work

2.1 Pixel Level-based Representation

The early work encode the image into an vector of fixed length as the representation for image captioning [Vinyals *et al.*, 2015]. The major issue caused by using this representation is its heavy compresses and mixture. Inspired by the CNNs in visual extraction for image classification, grid features generated by ResNets are used in captioners. For example, the pre-trained ResNet101 generates grid features and feeds them into Transformer [Vaswani *et al.*, 2017] to infer target words [Gao *et al.*, 2022]. Recently, ViT [Dosovitskiy *et al.*, 2020] and SwinT are used to extract the grid features to build one-stage image captioners. For example, to consider the semantic concepts, the VitCAP [Fang *et al.*, 2022] introduces the visual token to predict the semantic classification. However, ViTCAP still introduces other prior knowledge (i.e., multi-label classification as the concept information) to optimize the models. These mean that ViT and SwinT are promising to reduce the learning gap between vision and text. However, existing one-stage captioners embed the visual features based on the outputs with fixed sight, without adaption to the feature embedding of different distances.

2.2 Regional Level-based Representation

The detection-based methods extract the region-level features as the visual representation by the fixed detector. The salient objects of the image can be captured as a set of feature vectors, which greatly reduces the difficulty of visual semantic embedding and improves the performance of image captioning [Wang *et al.*, 2022a]. For example, up-down model [Anderson *et al.*, 2018] encodes the input image with a set of objects (i.e., RoI-pooled features) detected by a frozen Faster-RCNN [Ren *et al.*, 2015] pre-trained on Visual Genome [Krishna *et al.*, 2016]. To further compute the spatial geometry relationships [Hu *et al.*, 2018] for image captioning, Object Relation Transformer [Herdade *et al.*, 2019] explicitly incorporates the relative geometric position and size with the semantic relationships to enrich the embedded features. Conditioned on the region features, M^2 [Cornia *et al.*, 2020] infers the captions through learning a multi-level representation of the relationship between regions to exploit low- and high-level features. Note that those regional level-based methods learn prior knowledge based on the detection tags. So, the task-based information gap between the detection-centric features and the captioning-guided features makes these two-stage captioners suboptimal, which may decrease the captioning performance. Therefore, BPTOD [Kuo and Kira, 2022] mines attributes and relationships from the Visual Genome dataset as an auxiliary input to represent missing information to improve performance. However, BPTOD calculates both contextual descriptions and region features by using the multi-modal pre-trained model and frozen detector, respectively, which still is not an end-to-end trainable model.

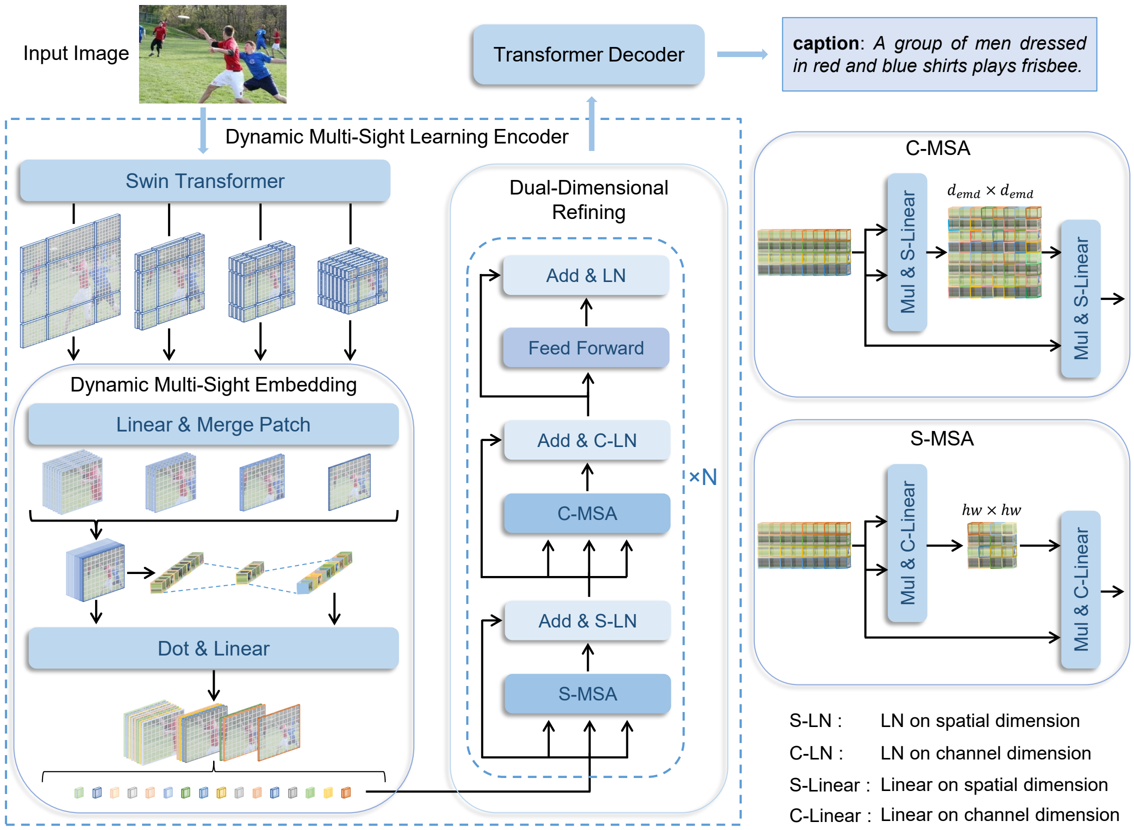

Figure 2: Overview of our OSIC framework, which includes two main components, i.e., standard Transformer decoder and a dynamic multi-sight learning encoder that consists of SwinT equipped with dynamic multi-sight embedding and dual-dimensional refining.

3 Proposed Method

As shown in Figure 2, our OSIC includes a standard Transformer decoder and a dynamic multi-sight learning encoder, i.e., SwinT equipped with a newly proposed dynamic multi-sight embedding and a cascaded dual-dimensional refining.

3.1 Captioning Procedure

Conditioned an input image I, our OSIC infers and generates a descriptive sentence S. Firstly, multi-level grid features $G = \{g_i\}$,($i = 1, 2, 3, 4$) are learnt by SwinT from I. Then, a linear embedding layer followed by patch-merging further extracts the multi-sights features M from the G. Specifically, M is the concatenation of grid features $\{g_i\}$. After that, salience coefficients E of the grid features of different sights are calculated by DMSE through the average pooling and linear projections to squeeze and dynamically excite the salient features in relevant sights. The output of the DMSE is tiled into a feature sequence as the input of the DDR, which consists of N operation layers, performing non-local interaction in spatial and channel dimensions. Each DDR layer is followed by a feed-forward network [Vaswani *et al.*, 2017] separately. Finally, the refined features are decoded by a Transformer decoder to generate the descriptive sentence S.

3.2 Dynamic Multi-Sight Embedding (DMSE)

Due to missing the visual details, it is inadequate to use only the representation of global compressed (i.e., the pooling features in the last layer of CNNs) for the input image, while the multi-sights of grid features can have more local details benefitting the captioning. Simultaneously, simply merging the grid features of multi-sights may confuse the visual embedding. Thus, the proposed DMSE first calculates the salience coefficients of the multi-sights features by linear projecting the global pooling of all sights. By considering the importance of global optimization for image captioning, we obtain a group of learnable coefficients based on the global linear projections of multi-sights. Specifically, we squeeze the grid features in each channel by average pooling to obtain a representative value sequence V_s, whose length is equal to the number of feature channels. Then, we connect the sequence V_s with the salience sequence by two layers of linear connections, so the salience coefficients E of multi-sights for subsequent captioning can be generated formally as:

$$E = L_c^{d_c}\{L_{d_c}^{c}[(L_{Hd_v}^{1}\langle M^T\rangle)^T]\}, \tag{1}$$

where M denotes the concatenation of grid features $\{g_i\}$, $(\cdot)^T$ denotes the transpose operation, and $L_i^j(\cdot)$ denotes the linear projection to map a tensor with embedding size i into that of j, c is the number of channel dimension. d_c is the number of the channel dimension of the squeezed features.

Then, the output of the DMSE module is obtained by multiplying the salience coefficients E by M as follows:

$$M_e = Norm_l(M \cdot E) + M, \tag{2}$$

where $Norm_l(\cdot)$ denotes the layer normalization, which is followed by a shortcut operation. After that, the output of DMSE are fed into the following DRR for further processing.

3.3 Dual-Dimensional Refining (DDR)

Given the embedded features M_e from the DMSE, we further feed them into the DRR. To improve the global-representational ability, we further refine the M_e via building non-local information interaction in both spatial and channel dimensions. Each non-local interaction of the two dimensions is modeled by computing the scaled dot product of them.

The layer normalization is operated at the corresponding dimension in which the dependence of pixels is computed. The processing output of the spatial position dimension or channel dimension in the DDR layer is obtained as follows:

$$M_r^i = Norm_l^i \left(\frac{M_e^Q \cdot \left(M_e^K\right)^T}{\sqrt{d_i}} \cdot M_e^V \right) + M_e, \tag{3}$$

where i denotes the interaction in either spatial or channel dimension, $Norm_l^i(\cdot)$ is the layer normalization operated in i-th dimension. Then, the parallel refining can be formulated as follows:

$$M_r^{pa} = M_r^s + M_r^c, \tag{4}$$

where M_r^s denotes the output from the refining layer of single-spatial dimension, and M_r^c denotes the output of the refining layer of single-channel dimension. Furthermore, the cascade refining calculates features as follows:

$$M_r^{ca} = Norm_l^c \left(\frac{M_r^{sQ} \cdot \left(M_r^{sK}\right)^T}{\sqrt{d_c}} \cdot M_r^{sV} \right) + M_r^s, \tag{5}$$

where $Norm_l^c(\cdot)$ denotes the layer normalization operated in channel dimension, M_r^{sQ}, M_r^{sK}, and M_r^{sV} are the linear projection representations of the outputs from the non-local refining layers of the multi-head self-attention on a spatial dimension, respectively, and d_c is the length of the bottom row vector $M_r^{sK}(Hd_v, :)$ of $M_r^{sK} \in R^{Hd_v \times d_m}$, i.e., d_m.

After that, the refined grid features M_r^i are fed into the feed-forward network and sequentially processed by repeating N times the above operation layer (where N is the number of layers). The refined features are finally decoded by a standard Transformer decoder to generate sentences.

3.4 Objective Function

We use two objective functions for optimization in the training process, following the widely used benchmarks. It consist of the cross-entropy loss (XE) for the maximum log-likelihood training and the reinforcement learning loss using the CIDEr score as a reward for self-critical training (SC) [Rennie *et al.*, 2017]. For XE training, with respect to the parameters θ and ground truth sentence $y^*_{(1:T)}$, the XE loss is calculated for the optimization as follows:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log \left(p\left(y_t^* | y_{0:t-1}^*, I, \theta \right) \right). \tag{6}$$

For the SC training, the model is fine-tuned continually by optimizing the non-differentiable CIDEr score as the reward of reinforcement learning processing formally as:

$$\nabla_\theta L_{SC}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left[r\left(y_{1:T}^i \right) - b \right] \nabla_\theta \log p_\theta \left(y_{1:T}^i \right), \tag{7}$$

where n denotes the beam size, $r(\cdot)$ denotes the CIDEr-D score function, and $b = (\sum_i r(y_{1:T}^i))/n$ is the greedily decoded score value generated by the current model.

4 Experiments

4.1 Experimental Settings

Datasets. The experiments are mainly conducted on the MSCOCO [Lin *et al.*, 2014] dataset, and further generalized on Flickr8K [Hodosh *et al.*, 2013] and Flickr30K [Young *et al.*, 2014] datasets. MSCOCO is the most widely-used and competitive benchmark, which has 164,062 images annotated with 5 ground truth captions for each image. For the fair comparison, we follow the Karpathy's splits [Andrej Karpathy, 2017], in which MSCOCO dataset consists of 113,287 training images, 5,000 validation images, 5,000 offline testing images and 40,775 online testing images with 5 and 40 human-annotations. For training, validation and testing, the Flickr8K dataset containing 8,091 images is split into 6,091 images, 1000 images and 1000 images respectively, and the Flickr30K dataset (31,014 images) is split into 29,000 images, 1014 images and 1000 images respectively.

Evaluation Metrics. The generated sentences are fairly evaluated by using the widely-used metrics, i.e., BLEU-1/4 [Papineni *et al.*, 2002], CIDEr [Vedantam *et al.*, 2015], METEOR [Banerjee and Lavie, 2005] and ROUGE-L [Lin, 2004]. They are denoted as B-1/4, C, M, and R for short.

Implementation Details. For training, we first train our model under XE loss for 15 epochs with a mini-batch size of 8, and an Adam optimizer whose learning rate is initialized at 4e-4 with the warmup-step of 20,000. The learning rate is decayed 0.1 times from the 9-th epoch on. We increase the scheduled sampling probability by 0.05 for every 3 epochs. After the XE training, we train our model by optimizing the CIDEr score with the SC training strategy for another 15 epochs with an initial learning rate of 4e-5, which is decayed 0.1 times every 4 epochs. For testing, we use the beam search for our model with a beam size of 2. The default random seed is set to 42. All experiments are conducted in a single NVIDIA RTX2080Ti GPU with Pytorch 1.7 platform.

4.2 Main Results

Offline Evaluation. On the offline MSCOCO Karpathy's test, we show the evaluation of each method in Table 1. The comparied methods can be roughly divided into two groups:

Methods	Cross-Entropy Loss							CIDEr Score Optimization						
	B@1	B@2	B@3	B@4	M	R	C	B@1	B@2	B@3	B@4	M	R	C
Two-Stage Methods														
RFNet (*iccv2018*)	76.4	60.4	46.6	35.8	27.4	56.5	112.5	79.1	63.1	48.4	36.5	27.7	57.3	121.9
Up-Down (*cvpr2018*)	77.2	–	–	36.2	27.0	56.4	113.5	79.8	–	–	36.3	27.7	56.9	120.1
ORT (*nips2019*)	76.6	–	–	35.5	28.0	56.6	115.4	80.5	–	–	38.6	28.7	58.4	128.3
AoANet (*iccv2019*)	77.4	–	–	37.2	28.4	57.5	119.8	80.2	–	–	38.9	29.2	58.8	129.8
M^2T (*cvpr2020*)	–	–	–	–	–	–	–	80.8	–	–	39.1	29.2	58.6	131.2
X-Transformer (*cvpr2020*)	77.3	61.5	47.8	37.0	28.7	57.5	120.0	80.9	65.8	51.1	39.7	29.5	59.1	132.8
DRT (*acm mm2021*)	–	–	–	–	–	–	–	81.7	–	–	40.4	29.5	59.3	133.2
RSTNet (*cvpr2021*)	–	–	–	–	–	–	–	81.8	–	–	40.1	29.8	59.5	135.6
S^2 Transformer (*ijcai2022*)	–	–	–	–	–	–	–	81.1			39.6	29.6	59.1	133.5
FutureCap (*acm mm2022*)	–	–	–	–	–	58.2	–	**82.2**	–	–	40.3	30.1	59.8	136.3
LightCap (*aaai2023*)	–	–	–	–	–	–	–	–	–	–	40.1	29.9	–	136.6
ConCap (*aaai2023*)	–	–	–	–	–	–	–	–	–	–	40.5	**30.9**	–	133.7
SCD-Net (*cvpr2023*)	**79.0**	**63.4**	**49.1**	37.3	28.1	58.0	118.0	81.3	66.1	51.1	39.4	29.2	59.1	131.6
One-Stage Methods														
PTSN (*acm mm2022*)	–	–	–	–	–	–	–	81.7	–	–	39.7	29.5	58.6	134.7
BPTOD (*cvpr2022*)	–	–	–	–	–	–	–	81.5	–	–	39.7	30.0	59.5	135.9
ViTCAP (*cvpr2022*)	–	–	–	35.7	28.8	57.6	121.8	–	–	–	40.1	29.4	59.4	133.1
OSIC (ours)	78.5	62.8	**49.1**	**38.0**	**29.1**	**58.3**	**124.2**	**82.2**	**67.2**	**53.0**	**41.0**	29.8	**60.2**	**137.2**

Table 1: Performance (%) comparison on the MSCOCO Karpathy's test split.

i) Two-stage methods, which adopt offline features directly to infer descriptions, including RFNet [Jiang *et al.*, 2018], Up-Down [Anderson *et al.*, 2018], ORT [Herdade *et al.*, 2019], AoANet [Huang *et al.*, 2019], M^2T [Cornia *et al.*, 2020], X-Transformer [Pan *et al.*, 2020], DRT [Song *et al.*, 2021], RSTNet [Zhang *et al.*, 2021], S^2Transformer [Zeng *et al.*, 2022a], FutureCap [Fei *et al.*, 2022], LightCap [Wang *et al.*, 2023b], ConCap [Wang *et al.*, 2023a] and SCD-Net [Luo *et al.*, 2023]; ii) One-stage methods, which optimize features extracting and captioning simultaneously, including PTSN [Zeng *et al.*, 2022b], BPTOD [Kuo and Kira, 2022] and ViTCAP [Fang *et al.*, 2022]. Our OSIC belongs to the one-stage method.

We first compare our method with the two-stage methods in Table 1. By closing the task-based information gap, our OSIC gains with 41.0 (+1.2%) in B-4 and 137.2 (+0.44%) in CIDEr respectively. Our proposed OSIC compares favorably with all previous methods across almost all metrics. This proves the effectiveness of our proposed OSIC. Then, we compare our method with the one-stage models, including PTSN, BPTOD and ViTCAP. In spite of additional information used in these methods (e.g., retrieved text and image conditioning from pre-trained CLIP [Radford *et al.*, 2021] for BPTOD, and multi-label classification for ViTCAP), our OSIC achieves better performance by using the proposed DMSE and DDR. Note that our model is only trained on the image-text pairs, without other additional information, which has a lower annotation cost of the dataset than them. Moreover, ConCap and LightCap have introduced large vision and language models (LVLM), such as pretrained Clip and Bert [Devlin *et al.*, 2018], respectively. Compared with these LVLM-based captioners, our OSIC is still competitive.

GT1: a dog and a bike sitting next to a red wall
GT2: a dog sitting on a sidewalk with a bike near by
GT3: A bike and a dog on the sidewalk outside a red building
Baseline: a dog sitting on a bicycle next to a bicycle
OSIC: a dog standing next to a bike in front of a red building

GT1: a cat sleeping against a stuffed polar bear.
GT2: a black and white cat sleeps next to a stuffed bear.
GT3: a white teddy bear with a cat sleeping beside it.
Baseline: a stuffed bear sitting next to a teddy bear
OSIC: a black and white cat laying next to a stuffed teddy bear

GT1: A couple of small birds standing on top of a table.
GT2: Two birds perched on a table near a plate of food.
GT3: two little sparrows standing on a table by a knife
Baseline: a plate of food on a table
OSIC: two birds sitting on a table next to a bottle of wine

GT1: The motorcycle has a large cargo box attached to the seat.
GT2: A blue motorcycle with a luggage compartment parked at a driveway.
GT3: A motorcycle is parked on a sidewalk in front of a yard
Baseline: a motorcycle parked on a street
OSIC: a blue motorcycle parked on the side of a street

Figure 3: Visualization of some captioning examples

Online Evaluation. We further evaluate our OSIC on the official COCO test by submitting our generated captions to the online test server [*] in Table 2. It is noteworthy that the performances of our OSIC are reported with a single model, without using any ensemble models. From the observations, our OSIC again surpasses all the single models across all metrics. Moreover, the single model of our OSIC even attains

[*]https://competitions.codalab.org/competitions/3221

Methods	B@1		B@2		B@3		B@4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Ensemble Model														
GCN-LSTM [Yao *et al.*, 2018]	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
AoANet [Huang *et al.*, 2019]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
XT (Res-101) [Pan *et al.*, 2020]	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
M^2T [Cornia *et al.*, 2020]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
CtxAdpAtt [Wang *et al.*, 2022b]	81.0	95.2	65.5	91.0	51.5	81.7	39.3	70.9	29.4	39.0	59.6	75.1	128.5	131.0
UAIC [Fei *et al.*, 2023]	81.9	96.3	66.5	91.1	51.8	83.0	39.6	72.9	29.2	38.9	59.2	74.7	129.0	132.8
Single Model														
CAVP [Liu *et al.*, 2018]	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
SGAE [Yang *et al.*, 2019]	80.6	95.0	65.0	88.9	50.1	79.6	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
CMAL [Guo *et al.*, 2021]	79.8	94.3	63.8	87.2	48.8	77.2	36.8	66.1	27.9	36.4	57.6	72.0	119.3	121.2
SAIC [Yan *et al.*, 2021]	80.0	94.5	64.1	88.2	49.2	78.8	37.2	67.8	28.0	36.8	57.7	72.4	121.4	123.7
VASS [Wei *et al.*, 2021a]	79.9	94.7	64.6	88.8	50.0	79.9	38.0	69.4	27.9	37.0	58.1	73.3	120.7	123.2
SCD-Net [Luo *et al.*, 2023]	80.2	95.1	64.9	89.3	50.1	80.1	38.1	69.4	29.0	38.2	58.5	73.5	126.2	129.2
OSIC (ours)	**81.6**	**95.7**	**66.6**	**90.6**	**52.0**	**82.1**	**39.9**	**71.9**	**29.3**	**38.8**	**59.4**	**74.7**	**130.5**	**133.1**

Table 2: Leaderboard of various captioning models on the online MS COCO test server.

Methods	Flickr8k							Flickr30k						
	B@1	B@2	B@3	B@4	M	R	C	B@1	B@2	B@3	B@4	M	R	C
ATT-FCN [You *et al.*,][2016]	–	–	–	–	–	–	–	64.7	46.0	32.4	23.0	18.9	–	–
SCA-CNN [Chen *et al.*, 2017]	68.2	49.6	35.9	25.8	22.4	–	–	66.2	46.8	32.5	22.3	19.5	–	–
Adaptive [Lu *et al.*, 2017]	–	–	–	–	–	–	–	67.7	49.4	35.4	25.1	20.4	–	53.1
DA [Gao *et al.*, 2019]	–	–	–	–	–	–	–	73.8	55.1	40.3	29.4	23.0	–	66.6
SDCD [Ding *et al.*, 2020]	67.2	45.1	30.5	21.5	–	–	–	66.3	43.7	29.2	21.1	–	–	–
G-NIC+P+D Att [Yu *et al.*, 2021]	68.4	50.3	37.0	22.8	22.6	–	–	69.7	46.1	35.5	23.7	20.4	–	–
VASS [Wei *et al.*, 2021a]	–	–	–	–	–	–	–	73.2	56.0	41.5	30.6	22.7	50.8	66.0
OSIC (ours)	**73.8**	**56.7**	**41.9**	**30.2**	**24.9**	**53.7**	**82.9**	**76.8**	**59.4**	**44.9**	**33.6**	**24.8**	**54.0**	**83.6**

Table 3: Performance (%) comparison on the Flickr8K and Flickr30K.

competitive performances with the ensemble version of some captioners (e.g., GET and UAIC), which proves the advantage of our proposed one-stage captioner.

Qualitative Results. We show some examples predicted by the baseline and our OSIC in Figure 3. In this paper, the baseline is set as extracting visual features only from the last layer of the SwinT, followed by the standard Transformer decoder. Clearly, our OSIC catches additional details and generates more descriptions with accurate semantic relationships, since our OSIC model can directly and dynamically embed more visual sources of multi-sights. For example, our model accurately captures the key scene in the top picture (i.e., "in front of a red building"), while the baseline misses it.

Generalization on other datasets. Recently, the captioners tend to be rarely evaluated on the Flickr8K and Flickr30K, since that: 1) The number of images in them is much less than the MSCOCO; 2) Their images focus on human activities, which makes the captioner difficult to describe images on large scale scenes. However, it still remains an effective benchmark for evaluating the generalization of models. As shown in Table 3, our OSIC achieves the best results on all standard metrics, and outperforms the current models by a large margin, which proves its well-generalization.

4.3 Ablation Study

Ablation on single DMSE and DDR. From the first and second rows of Table 4, as adding the DMSE into the baseline, the performance is greatly improved across all metrics over the baseline, which proves the effectiveness of the proposed DMSE. As shown in the 3rd to 6-th rows in Table 4, we study the effectiveness of the DDR, which contains the non-local interactions in spatial/channel dimension, or combines them in a parallel/cascade mode, respectively. Clearly, our OSIC largely benefits from refining the features on spatial or channel dimension, with improvements of at least $+11.8\%$ in Bleu-1 and $+29.3\%$ in CIDEr over baseline. All the above demonstrates that the performance gain indeed comes from the DMSE and DDR, which largely raised the metrics.

Ablation on joint DMSE and DDR. We incorporate the DMSE and the DDR with four kinds of non-local modes into the baseline, as shown in the 7-th to 10-th rows in Table 4. Clearly, OSIC with both DMSE and DDR can further deliver better results. Especially, cascaded non-local interactions in spatial and channel dimensions generally perform the best.

Baseline	Multi-sight embedding	Feature refining				B@1	B@2	B@3	B@4	M	R	C
		Spatial	Channel	Parallel	Cascade							
✓						68.8	53.8	40.3	29.4	23.6	53.1	92.2
✓	✓					77.9	62.1	48.2	37.2	28.8	57.8	122.2
✓		✓				76.9	60.9	47.2	36.3	28.5	57.0	119.2
✓			✓			77.2	61.1	47.2	36.2	28.5	57.2	119.3
✓				✓		77.5	61.5	47.6	36.6	28.5	57.4	119.9
✓					✓	77.5	61.6	47.8	36.8	28.6	57.4	120.9
✓	✓	✓				78.0	62.4	48.6	37.6	29.0	58.0	123.1
✓	✓		✓			78.2	62.4	48.6	37.5	28.9	58.0	122.6
✓	✓			✓		**78.6**	**63.0**	**49.1**	**38.0**	28.9	58.2	123.1
✓	✓				✓	78.5	62.8	**49.1**	**38.0**	**29.1**	**58.3**	**124.2**

Table 4: Ablation study on the importance of each module. DDR is further ablated as four modes: singly refining in spatial/channel dimension, and combining them in parallel/cascade way. The result is obtained by XE Loss training on the MSCOCO Karpathy's test split.

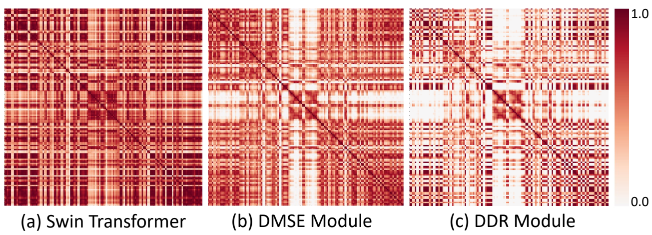

Figure 4: Heatmaps of the correlations between features at different levels. The heavier the color, the closer the pairwise relationship.

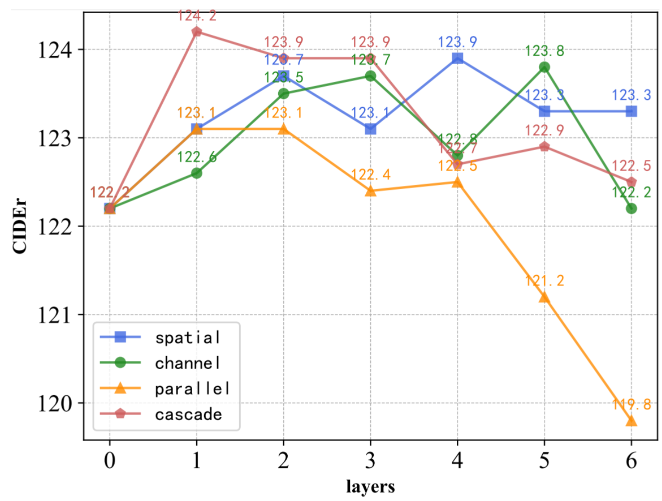

Figure 5: CIDEr of different refinings, various numbers of layers.

Since the slight outperformance, our model in a cascade mode is illustrated in Figure 2. We also visualize the heatmaps of correlations between features to analyze the effectiveness of DMSE and DDR in Figure 4. Specifically, DMSE builds relatively explicit independence of features. DDR further refines the relationship among features globally. The sparsification in heatmaps denotes few correlations among features on the global scale. As additional modules are added, the value in dense regions denotes higher connections with other features, while that in sparse regions denotes fewer connections. For example, for the local features in the top left corner in Figure 4, DMSE supplements and enriches the originally missed connections, and also weakens the unnecessary connections. Based on DMSE, the DDR further refines the embedded features. That is, important connections are strengthened and unimportant connections are weakened, without changing the overall feature distributions. From the above ablation studies, we conclude that both the DMSE and DDR are important to ensure the effectiveness of our OSIC for image captioning.

Impact of the DDR settings. We ablate our OSIC with different settings on the modes and the number of the DDR layer, as shown in Figure 5. We vary the number of refining layers from 0 to 6. From the observation, in parallel and cascade modes, the model with only 1 layer of refining can perform better. We see that parallel refining generally performs the worst among different modes. It can also be seen that our OSIC with only 1 layer cascade refining outperforms other cases. That is, cascading the spatial and channel dimensions is the best setting to obtain the best performance, which is used in all the experiments of this paper. It demonstrates that OSIC works better without needing lots of parameters, which benefits the fast inferring and text generating.

5 Conclusion

We first define the task-based information gap that exists in current two-stage captioners, and address it by presenting a novel one-stage image captioner called OSIC. OSIC directly captures the different sight of representations of the image by a new dynamic multi-sight learning encoder refined by a dual-dimensional refining, then decodes the features into captions. The visual representation is improved by building non-locally dual-dimensional interaction. Extensive simulations demonstrated the effectiveness of our OSIC attributed to the dynamic multi-sight embedding and dual-dimensional refining, in comparison to other related methods. We also conduct extensive ablation studies to explore the contribution of modules and settings. In the future, we will explore more efficient and robust image captioners under complex real-world conditions, such as describing the rainy images [Wei *et al.*, 2021b; Wei *et al.*, 2022] or blur scenes [Zhao *et al.*, 2022].

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62072151), Anhui Provincial Natural Science Fund for the Distinguished Young Scholars (2008085J30), Open Foundation of Yunnan Key Laboratory of Software Engineering (2023SE103), CCF-Baidu Open Fund and CAAI-Huawei MindSpore Open Fund.

References

[Anderson *et al.*, 2018] 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 *CVPR*, 2018.

[Andrej Karpathy, 2017] Li Fei-Fei Andrej Karpathy. Deep visual-semantic alignments for generating image descriptions. *TPAMI*, 2017.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, 2005.

[Chen *et al.*, 2017] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CORR*, 2018.

[Ding *et al.*, 2020] Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan. Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, 2020.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[Fang *et al.*, 2022] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *CVPR*, 2022.

[Fei *et al.*, 2022] Zhengcong Fei, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Efficient modeling of future context for image captioning. In *ACM MM*, 2022.

[Fei *et al.*, 2023] Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Uncertainty-aware image captioning. In *AAAI*, 2023.

[Gao *et al.*, 2019] Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. Deliberate attention networks for image captioning. In *AAAI*, 2019.

[Gao *et al.*, 2022] Yi-Meng Gao, Ning Wang, Wei Suo, Mengyang Sun, and Peifeng Wang. Improving image captioning via enhancing dual-side context awareness. In *ICMR*, 2022.

[Guo *et al.*, 2021] Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. In *IJCAI*, 2021.

[Herdade *et al.*, 2019] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, 2019.

[Hodosh *et al.*, 2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013.

[Hu *et al.*, 2018] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018.

[Huang *et al.*, 2019] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[Jiang *et al.*, 2018] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, W. Liu, and T. Zhang. Recurrent fusion network for image captioning. In *ICCV*, 2018.

[Krishna *et al.*, 2016] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016.

[Kuo and Kira, 2022] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *CVPR*, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

[Liu *et al.*, 2017] Chang Liu, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Loddon Yuille. Mat: A multimodal attentive translator for image captioning. In *IJCAI*, 2017.

[Liu *et al.*, 2018] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *ACM MM*, 2018.

[Lu *et al.*, 2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.

[Luo *et al.*, 2023] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *CVPR*, 2023.

[Pan *et al.*, 2020] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2015.

[Rennie *et al.*, 2017] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[Song *et al.*, 2021] Zeliang Song, Xiaofei Zhou, Linhua Dong, Jianlong Tan, and Li Guo. Direction relation transformer for image captioning. In *ACM MM*, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[Wang *et al.*, 2022a] Bo Wang, Zhao Zhang, Jicong Fan, Mingbo Zhao, Choujun Zhan, and Mingliang Xu. Fineformer: Fine-grained adaptive object transformer for image captioning. In *ICDM*, 2022.

[Wang *et al.*, 2022b] Yiyu Wang, Jungang Xu, and Yingfei Sun. A visual persistence model for image captioning. *Neurocomputing*, 2022.

[Wang *et al.*, 2023a] Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. Controllable image captioning via prompting. In *AAAI*, 2023.

[Wang *et al.*, 2023b] Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, and Linlin Li. Efficient image captioning for edge devices. In *AAAI*, 2023.

[Wei *et al.*, 2021a] Haiyang Wei, Zhixin Li, Feicheng Huang, Canlong Zhang, Huifang Ma, and Zhongzhi Shi. Integrating scene semantic knowledge into image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2021.

[Wei *et al.*, 2021b] Yanyan Wei, Zhao Zhang, Yang Wang, Mingliang Xu, Yi Yang, Shuicheng Yan, and Meng Wang. Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking. *TIP*, 2021.

[Wei *et al.*, 2022] Yanyan Wei, Zhao Zhang, Huan Zheng, Richang Hong, Yi Yang, and Meng Wang. Sginet: Toward sufficient interaction between single image deraining and semantic segmentation. In *ACM MM*, 2022.

[Yan *et al.*, 2021] Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. Semi-autoregressive image captioning. In *ACM MM*, 2021.

[Yang *et al.*, 2019] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[Yao *et al.*,] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *CVPR*.

[Yao *et al.*, 2018] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[You *et al.*,] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014.

[Yu *et al.*, 2021] Litao Yu, Jian Zhang, and Qiang Wu. Dual attention on pyramid feature maps for image captioning. *TMM*, 2021.

[Zeng *et al.*, 2022a] Pengpeng Zeng, Haonan Zhang, Jingkuan Song, and Lianli Gao. S2 transformer for image captioning. In *IJCAI*, 2022.

[Zeng *et al.*, 2022b] Pengpeng Zeng, Jinkuan Zhu, Jingkuan Song, and Lianli Gao. Progressive tree-structured prototype network for end-to-end image captioning. In *ACM MM*, 2022.

[Zhang *et al.*, 2021] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *CVPR*, 2021.

[Zhao *et al.*, 2022] Suiyi Zhao, Zhao Zhang, Richang Hong, Mingliang Xu, Yi Yang, and Meng Wang. Fcl-gan: A lightweight and real-time baseline for unsupervised blind image deblurring. In *ACM MM*, 2022.