Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning

Xinzhi Dong
School of Computer, Wuhan University
Wuhan, Hubei, China
dongxz97@whu.edu.cn

Chengjiang Long*
JD Finance America Corporation
Mountain View, CA, USA
cjfykx@gmail.com

Wenju Xu
OPPO US Research Center, InnoPeak Technology Inc.
Palo Alto, CA, USA
xuwenju123@gmail.com

Chunxia Xiao†
School of Computer, Wuhan University
Wuhan, Hubei, China
cxxiao@whu.edu.cn

ABSTRACT

Existing image captioning methods just focus on understanding the relationship between objects or instances in a single image, without exploring the contextual correlation existed among contextual image. In this paper, we propose Dual Graph Convolutional Networks (Dual-GCN) with transformer and curriculum learning for image captioning. In particular, we not only use an object-level GCN to capture the object to object spatial relation within a single image, but also adopt an image-level GCN to capture the feature information provided by similar images. With the well-designed Dual-GCN, we can make the linguistic transformer better understand the relationship between different objects in a single image and make full use of similar images as auxiliary information to generate a reasonable caption description for a single image. Meanwhile, with a cross-review strategy introduced to determine difficulty levels, we adopt curriculum learning as the training strategy to increase the robustness and generalization of our proposed model. We conduct extensive experiments on the large-scale MS COCO dataset, and the experimental results powerfully demonstrate that our proposed method outperforms recent state-of-the-art approaches. It achieves a BLEU-1 score of 82.2 and a BLEU-2 score of 67.6. Our source code is available at *https://github.com/Unbear430/DGCN-for-image-captioning*.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Scene understanding**.

KEYWORDS

Image Captioning; Transformer; Graph Neural Network; Curriculum Learning

---

*This work was co-supervised by Chengjiang Long and Chunxia Xiao.
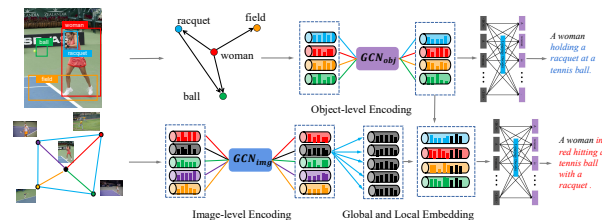†Corresponding author.

Figure 1: An illustration example of image captioning task, taking Graph Convolutional Networks (GCNs) as the visual encoder and a sequence-to-sequence model as a linguistic decoder. The top represents the visual encoder via a single object-level GCN on a single image, and the bottom indicates the visual encoder through the combination of both the object-level GCN and an image-level GCN. The visual embedding feature extracted in either way can be fed into a linguistic decoder to generate a reasonable text description.

1 INTRODUCTION

Image captioning aims to generate a syntactic and correct description for a given single image, *a.k.a.*, a translation from image to language. One example is illustrated in Figure 1. It requires a clear understanding of different objects and the relation between them, and the grammar in the caption. Otherwise, the model will miss some objects or generate corrections that do not conform to common sense. Many deep learning-based methods and algorithms have been proposed for image captioning [9, 24, 41, 46]. They heavily rely on Convolutional Neural Networks (CNNs) [22, 32, 37, 56, 59, 60] to extract hierarchical image features and take Recurrent Neural Networks (RNNs) like long-short-term memory (LSTM) [4, 25, 33, 45, 50] as the language model to generate descriptive captions. However, these CNN-RNN based methods suffer from the limitations of fixed

captions. Therefore, captions generated by these methods are far from being able to exactly describe the content of the input image.

Graph Convolutional Networks (GCNs) [42, 52, 55, 56] are then exploited on the structured graphs to enrich region representations for image captioning tasks. Recently, transformer [8, 43] has shown promising performance when dealing with serialization information. What's more, transformer has been recognized as the state of the art in sequence modeling tasks like language understanding and machine translation. Although there exists some efforts [5, 15] of applying transformer to solve the image captioning task and achieve some success, the potential of integrating GCNs with transformer is still largely under-explored.

In this paper, we propose a novel image captioning framework taking Dual Graph Convolutional networks (Dual-GCN) as the visual encoder and take the transformer as the linguistic decoder. As shown in Figure 2, our framework integrates an object-level GCN and an image-level GCN for visual feature encoding. The first GCN focuses on objects in local regions to explore spatial object-level relation [6, 39, 53, 61], while the second one targets on the image-level similarity relation [12, 47] amongst multiple similar images. We calculate the similarity of different images and select the images with high similarity as a meaningful complementary global visual embedding to ensure a more reasonable and accurate text description generation. In this way, our framework can learn the global and local visual embedding with the well-designed Dual-GCN framework. This significantly distinguishes our method from other GCN-based models [55, 56], which only considers the local object regions.

Unlike most existed approaches taking RNNs as the linguistic decoders, we choose transformer as our linguistic decoder instead. Due to the well-designed multi-head self-attentions, the transformer has the advantage of parallelization and has a stronger ability to extract visual features. It shows superior performance in the fusion of visual features and text features. Therefore, taking in the global and local visual embedding as input, our transformer linguistic decoder can generate a syntactic and correct description for the single input image.

It is worth mentioning that existing caption models [1, 5, 9, 21, 21, 45, 50, 52, 54–58] are trained in a straightforward manner, in which all the training data are fed into the model evenly and equally, ignoring the fact that different training images may contain information at different levels. We argue that a difficult image capturing a complex scene usually needs a long caption to describe the objects and their spatial positioning. Therefore, an under-trained model could be misled by the wrong gradients when trained on difficult images. Inspired by the human nature learning process in an easy-to-hard manner, we adopt the curriculum learning [49] as a training strategy. A cross-review mechanism is then introduced to distinguish the difficulty of the training images. In particular, we partition the entire training dataset evenly into multiple subsets and train an image caption model on each subset. For any training example, its difficulty level is determined across all the other learned models, each of which never takes it as the training example. With the difficulty levels assigned on all the training examples, we can train our proposed model on easy images at the beginning. Then we keep adding difficult images as the training examples. Obviously, the newly added information dramatically enhances our model's

robustness and generalization ability in terms of understanding different objects and the relation between them. To summarize, the main contributions of this paper are three-fold:

- We propose a novel framework that takes Dual-GCN as a powerful visual encoder and transformer as the linguistic decoder for image captioning. The first object-level GCN captures the region to region relation in a single image, and the second image-level GCN captures the image's similarity relation. This treatment enables both global and local visual information well encoded and fed into the transformer to generate better captioning descriptions.
- To our best knowledge, we are the first one to apply the Curriculum Learning as the training strategy on image captioning model in an easy-to-hard manner. Note that the difficulty level is determined by our well-designed cross-review mechanism.
- We conduct extensive experiments on the MS COCO dataset. The results strongly demonstrate that our proposed approach outperforms state-of-the-art methods. It achieves a BLEU-1 score of 82.2 and a BLEU-2 score of 67.6.

2 RELATED WORK

Image Captioning is a cross-modal task combining computer vision and natural language processing. It has been widely studied as a deep learning task [28] in recent years. The dominant paradigm in modern image captioning are sequence learning methods [9, 45, 50, 54, 57, 58]. Vinyals et al. [45] used a deep CNN to extract the visual features of the image and adopted LSTM [55] as the language model to guide the language sequence generation. Instead of learning a visual attention map on the input image, [38] proposed an encoder-decoder model that learns an adaptive attention map to combine the visual feature and language embedding. [1] proposed bottom-up attention and a top-down attention mechanism for caption generation. Chen et al. [3] extended the spatial -wise and channel-wise attention mechanism to extract both spatial features and semantic features from an image. Unlike these mentioned models, Yao et al. [55] adopted the graph neural network to integrate spatial and semantic relationships between different objects. In recent years, transformer-based methods [52] show great potential in image captioning. Herdade et al. [16] used the transformer architecture for image captioning. Then Li et al. [29] proposed to take an external tagger providing visual information and semantic knowledge to enhance the performance of the transformer model. As reverse process of image captioning, Hu et al. [19] [18] used diverse conditional GAN [51] for text representation and generated video animation from single still image [17]. Fang et al. [10] solved a series of problems of analyzing complex text and finally generating corresponding images, and also used scene graph and generates more realistic images [11].

Graph Neural Networks have been widely used in many applications to build up the relations between different features, such as nature language processing and computer vision tasks [20, 26, 34–36, 55]. In particular, [26] solves the computing bottleneck by learning the Laplacian polynomials of graphs. When combining with the LSTM [55], GCN can be used for image captioning task. It is used to extract relations between different objects, which achieved great
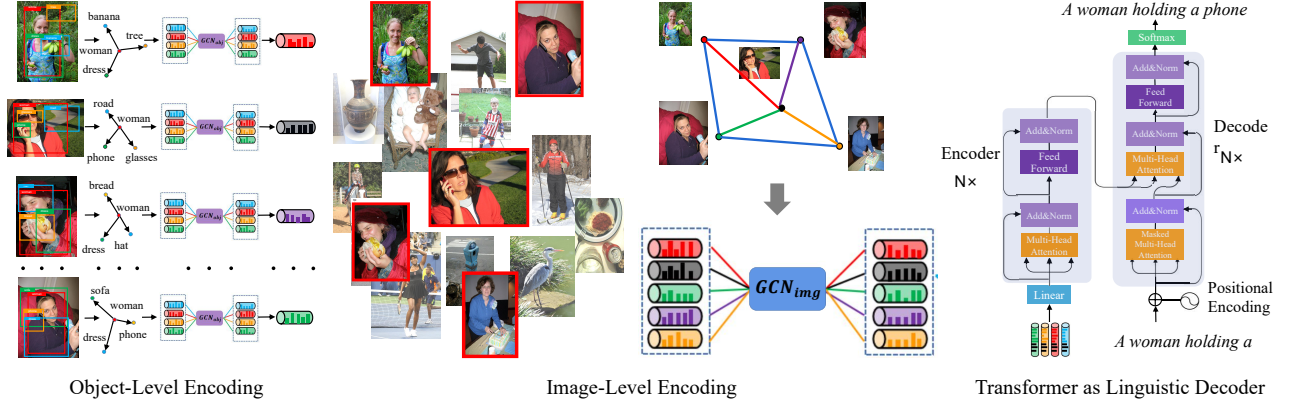
Figure 2: The overview of our proposed framework for image captioning. It consists of three main parts: an object-level GCN, an image-level GCN, and the transformer as the linguistic decoder. In specific, given an input image, we use an object-level GCN and an image-level GCN to re-encode both the object-level and image-level features. With a combination of visual embedding features as the input, we take transformer as the linguistic decoder to generate a reasonable text description.

results in the final caption generation. However, these GCN based methods only explore the object relations within one single image. Instead, we propose a Dual-GCN module to aggregate information extracted from both the input image and other similar images.

Transformer is first proposed by [43]. It is a combination of multi-layer encoder and decoder. The encoder consists of a stack of self-attention and feedforward layers, while the decoder uses self-attention on the text and the cross-attention to the last encoder layer. Herdade et.al [16] then used the transformer in image captioning. They combined geometric relationships between different input objects. [5] used a priori knowledge module to improve the transformer encoder and also used mesh connection between multi-layer encoder layer and decoder layer to improve the hierarchical diversity of feature extraction. Li et al. [29] used the transformer to make use of visual information and additional semantic knowledge given by an external tagger. In this paper, we take transformer as the linguistic model to generate a reasonable text description based on the extracted visual embedding feature.

Curriculum Learning is first proposed by [2] in the machine learning area. This method first classifies the sub-dataset according to the difficulty, such as finding examples that are easy to implement in the data set. Then it sets up an easy-to-hard curriculum for the learning procedure. Such a human-like learning strategy considerably improves the model's capability. Curriculum learning has been widely used in the computer vision area [13, 23, 48]. With the continuous expansion of transformer in computer vision, we propose to use a curriculum learning strategy aiming at improving our Dual-GCN model's performance on image captioning task.

3 PROPOSED METHOD

Our method aims to generate a caption that correctly describes the content of an input image. It comprises three main parts: Dual-GCN encoder, transformer decoder, and curriculum learning strategy. In this section, we will introduce each framework and our algorithm in detail. Figure 2 shows our main framework.

Given an image I, we take a textual caption to describe its content, denoted by S. First, we take the Faster-RCNN [41] as a backbone to produce a set of detected objects $V_{obj} = \{v_i\}_{i=1}^{O}$ where O denotes different object regions in an image. v_i denotes a C-dimensional feature representation of the i-th object region.

3.1 Dual-GCN Encoder

To learn a better latent representation of the input image, we propose a Dual-GCN encoder, which consists of an object-level GCN and a image-level GCN. Based on the region feature from the Faster-RCNN detector, our Dual-GCN encoder produces better features by seeking the spatial relation within the same image and the similarity among different images.

Object-Level GCN. We take the object-level GCN to build a spatial relation between different objects within the same image. To this end, we consider each feature extracted from one single object region as a vertex and consider the relative position amongst different objects as a directed line segment. Specifically, we build a spatial graph $G_{obj} = (V_{obj}, E_{obj})$, where $E_{obj} = \{(v_i, v_j)\}$ is the set of spatial relation edges between region vertices, and the edge (v_i, v_j) represents the relative geometry relationship between the i-th and the j-th objects. More specifically, following [55], we use the IOU, relative distance, and the relative angle between two objects as the relative geometry relationship to determine the edges, these relationships can be divided into "inside", "cover", "overlap" and "no" relation" and so on. Therefore, each vertex v_i is encoded via a modified GCN as:

$$v_i^{obj} = \sigma(\sum_{v_j \in N(v_i)} W_{v_j \to v_i} v_j + b_{v_j \to v_i}), \qquad (1)$$

where each v_i is a real-valued vector and it represents the information of the nodes in the graph. σ is an activation function, and $N(v_i)$ denotes a set of neighbor vertices of v_i. $v_j \to v_i$ indicates the direction from v_j to v_i. $W_{v_j \to v_i}$ and $b_{v_j \to v_i}$ denote the transformation matrix and the bias vector, respectively.

Image-Level GCN. The information extracted from a single image is very limited. As a remedy, we propose a novel image-level GCN, which enhances the information flow in the graph by considering other relevant images. In image feature extraction, we observe that

images sharing a certain degree of similarity to the input image can be an auxiliary information source for the input image captioning.

Let $\bar{v}_j$ stands for the latent representation of j-th image, which contains a set of objects $V_{obj}^j = \{v_i^{obj}\}_{i=1}^O$. We define:

$$\bar{v}_j = \frac{1}{O} \sum_{i=1}^O v_i^{obj}, \qquad (2)$$

Based on the latent representation, we seek similar images $\bar{v}_s \in \Omega(\bar{v}_j)$ according to the $\mathcal{L}_2$ distance between the features of the input image and other images from the same dataset. This is formulated as:

$$\bar{v}_s \in \Omega(\bar{v}_j) := \left[\sum_{t=1}^C (\bar{v}_s^t - \bar{v}_j^t)^2 \right]_K, \qquad (3)$$

where C is the dimension of the feature; $[\cdot]_K$ is to select K images with lowest distance score. Once we obtain the top-K images, we establish an image-level graph $G_{img} = (V_{img}, E_{img})$ by considering the image feature as a node V_{img} and the similarity between paired images as an edge E_{img}. Then we encode the features represented by each image node in the graph:

$$u_j^{img} = \sigma(\sum_{\bar{v}_s \in \Omega(\bar{v}_j)} W\bar{v}_s + b), \qquad (4)$$

where $\Omega(\bar{v}_j)$ denotes the set of similar images of $\bar{v}_j$. Note that in our image-level graph u_j^{img} represents the feature of an image and its K nearest neighbor images.

Global and Local Visual Embedding. Since we only need to generate the caption of a given image, we choose to extract the feature vector of GCN encoded in the node of the original image. Then we concatenate the feature vectors of GCN recoded on image-level similarity with the original image encoded on object-level as the input of a transformer as the language model for text prediction:

$$
\begin{aligned}
U_j &= Concat(V_{obj}^j, v_j^{img}) \\
&= [concat(v_1^{obj}, u_j^{img}), \ldots, concat(v_O^{obj}, u_j^{img})].
\end{aligned} \qquad (5)
$$

In this way, we combine the information extracted from the input image and other similar images. The final visual embedding representation integrates the characteristics of each object in the input image locally and the contextual object information from multiple similar images globally. Then, the global and local visual embedding information is fed into the linguistic decoder to ensure a reasonable caption generation.

3.2 Transformer as Linguistic Decoder

As shown in Figure 2, after we get the global and local embedding $U = \{U_j | j = 1, \ldots, O\}$ of the input image, we take a transformer decoder to generate a caption correctly describing the content of the image.

Encoding Layer. The encoding layer is based on multi-head attention layer, which can be formulated as:

$$U_{att} = softmax(\frac{W_q U (W_k U)^T}{\sqrt{d}}) W_v U, \qquad (6)$$

where W_q, W_k and W_v are the learnable parameters used to create the query, keys and values respectively; d is a constant used for

normalization. Besides, a residual structure is adopted to stabilize the attention module. This ensures that the transformer can extract the spatial information of different objects to a greater extent in the self-attention part. Thus, we achieve:

$$\bar{U} = AddNorm(U \oplus U_{att}), \qquad (7)$$

where $\oplus$ is an element-wise addition operation. Based on the above structure, multiple encoding layers are stacked in sequence so that each encoder layer receives the output calculated by the previous encoder layer. This is equivalent to creating a multi-layer high-level coding of the relationship between image regions, in which the higher coding layer can utilize and refine the previously recognized relationship and finally produce the output.

Decoding Layer. We take a similar multi-head attention module to process text and image feature.The decoder layer also contains feed-forward layer, and all components are encapsulated in addnorm operator. The structure of decoder layer can be described as:

$$Q = AddNorm(\bar{U}, \mathbf{Y}), \qquad (8)$$

where $\mathbf{Y}$ is the input sequence. Like the encoder layer, our decoder layer takes a similar self-attention and residual structure. And multiple decoder layers are also stacked in sequence. The purpose is to extract higher-level text and image features and generate more accurate word distribution probability:

$$\bar{Q} = AddNorm(Q \oplus Q_{att}), \qquad (9)$$

where Q_{att} stands for the output of a multi-head attention module in the decoding layer. Finally, by inputting the generated results into the softmax layer, the word probability distribution of the currently generated words is obtained:

$$P = Softmax(\bar{Q}). \qquad (10)$$

After getting the word probability distribution through the transformer, we get the probability distribution $P(S'|S, I)$ for the given image. Where S is the ground-truth caption embedding matrix, I is the image feature we used, and S' is the currently generated captions. Through the algorithm of conditional probability, our model gets the whole caption probability for the given image.

3.3 Curriculum Learning

To adopt the curriculum learning for our image captioning task, we have to determine the difficulty level from easy to hard.

Objective Function. Let D_{data} be the MSCOCO training dataset, and Θ be our Dual-GCN+Transformer caption model. In order to obtain a more comparable and stable difficulty score, we first uniformly split our training dataset D_{data} into M parts. Each sub-dataset is denoted by D_i ($i = 1, 2, 3...M$). Then we train our Dual-GCN+Transformer caption models Θ_i on them respectively. Each model can only use $1/K$ of the training dataset for training. The parameters of these models can be learned by the following optimization:

$$\Theta_i = \arg \min_{\Theta_i} \sum_{(I,S) \in D_i} \mathcal{L}_{ce}(\Theta_i(I), S), \qquad (11)$$
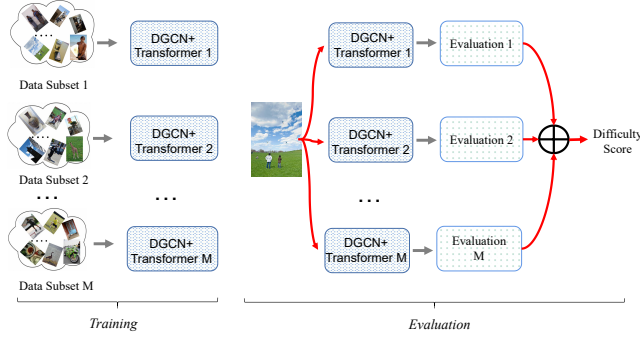
Figure 3: The cross-review mechanism used to determine the difficulty level for training examples. Inspired by [49], the dataset is evenly divided into M subsets. Then a Dual-GCN+Transformer model is trained on each data subset for caption generation. Given a training sample from the i-th data subset, we use all other $M - 1$ trained Dual-GCN+Transformer models together except the i-th model to determine its difficulty level. Note that we use BLEU as the metric to calculate the corresponding difficulty score.

where (I, S) is the image-to-text pair in the i-th traning dataset. $\mathcal{L}_{ce}$ defines the cross entropy loss function between the probability prediction by $\Theta_i(I)$ and the corresponding ground-truth caption S.

Difficulty Assessment. Most of the traditional ideas are based on word frequency. The difficulty is defined according to the number of times each word appears in the dataset. For example, in the MSCOCO training set, "person" appeared in 64,115 images and had 262,465 tag boxes, while "bus" only appeared in 3,952 images and had 6,069 tag boxes. Obviously, "person" is more accessible to generate than "bus". However, since image caption is a cross-modal matching task from image to text, the difficulty evaluation should be consistent with the corresponding evaluation metrics like BLEU-1 and BLEU-2. Inspired by [49], we adopt a cross review mechanism to determine the difficulty level for all the training examples, as shown in Figure 3.

After we train our caption models on M sub-datasets, we evaluate the difficulty level for each training example. Note that each image-to-text example $(I, S) \in D_i$ has been seen by model Θ_i during training. Hence we use another model Θ_k to evaluate the difficulty of (I, S):

$$\mathcal{E}_k(I, S) = 1 - Metric(\Theta_k(I), S), \qquad (12)$$

where $\mathcal{E}_k(I, S)$ is the difficulty score of the text-image pair (I, S). *Metric* stands for a formula, which could be one of the image captioning metrics such as BLEU-1,BLEU-2,BLEU-3 and BLEU-4 [40]. In this paper, we take the average of BLEU-1 score to indicate the difficulty. This is formulated as the sum of the evaluation scores of other caption models:

$$DS((I, S)) = \frac{1}{M-1} \sum_{(I,t) \in D_i, k \neq i} \mathcal{E}_k(I, S), \qquad (13)$$

where $DS((I, S))$ is the difficulty score of the text-image pair (I,S) in sub dataset D_i.

Curriculum Learning Arrangement Part. We first sort all text-image pairs according to their difficulty scores DS, and then divide them into M sub datasets: U_i. The training dataset is then arranged

Table 1: Visualization examples of images with different difficulty levels. We can observe that the difficulty level increases with more and more complex background.

Image I	Ground-truth Caption S	$DS(I,S)$
	A cheesy pizza siting on top of a table. *This large pizza has a lot cheese and tomato sauce.* *A prepared pizza being served at a table.* *A pizza with olive, cheese, tomato and onion.* *Baked pizza with herbs displayed on serving tray at table.*	0.214
	A group of Army soldiers visit a memorial. *A group of soldiers standing next to a bench.* *Soldiers performing a ceremony in a city park.* *Members of the military are in a park.* *Military officers standing on a public park in uniform.*	0.268
	Alittle in a pink shirt and khaki skirt playing tennis. *A young girl playing with her friend in a game of tennis.* *A girl is hitting a ball with a tennis racket.* *A small girl in action on a court with a racket.* *Military officers standing on a public park in uniform.*	0.308

from U_1 (the easiest) to U_M (the hardest). The numbers of example in each categories are defined as $|U_1|, |U_2|, \ldots, |U_M|$. We take M stages to train our model, it can be defined as C_i $(i = 1, \ldots, M)$. Notice that at each stage C_i, the image-to-text example is still shuffled to keep local stochastics, and each example from different stages does not overlap in order to prevent overfitting. For each learning stage C_i, we choose example according to a certain proportion from the aforementioned categories according to difficulty. The specific quantity is as follows:

$$\frac{|U_1|}{M} : \frac{|U_2|}{M} : \ldots \frac{|U_M|}{M}. \qquad (14)$$

When the training stages are reached on C_M, we think the model should be ready to train on the whole MS COCO dataset. So we add another learning stage C_{M+1}, which is the entire MS COCO dataset.

4 EXPERIMENTS

4.1 Datasets and Metrics

Microsoft COCO Dataset [31] (MS COCO) is the dataset most commonly used in image captioning tasks. This dataset consists of about 82,700 training images and 40,500 validation images. There are five human-annotated descriptions for each image. Following the settings used by [25], we take 5,000 images for validation, 5,000 images for testing and almost 110,000 images for training.

Visual Genome [27] is a large dataset containing almost 108,000 images with densely annotated objects, attributes and relationships. We take almost 100,000 images for training, 5,000 images for validation and 5,000 images for testing as done in [1] and [55]. We observe that there are approximately 51,000 images in the visual genome are also found in the MS COCO dataset. To avoid contamination of our MS COCO validation and test sets, we make sure that

Table 2: Comparison with the state of the art on the Microsoft COCO dataset, in single-model setting.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Up-down [1]	80.2	64.1	49.1	36.9	27.6	57.1	117.9
AOA-NET [21]	81.0	65.8	51.4	39.4	29.1	58.9	126.9
GCN-LSTM [55]	80.8	65.5	50.8	38.7	28.5	58.5	125.3
GCN-LSTM+HIP [56]	81.6	66.2	51.5	39.3	28.8	59.0	127.9
SGAE [52]	81.0	65.6	50.7	38.5	28.2	58.6	123.8
M^2 Transformer [5]	81.6	66.4	51.8	**39.7**	29.4	59.2	127.9
Dual-GCN+Transformer+CL	**82.2**	**67.6**	**52.4**	**39.7**	**29.7**	**59.7**	**129.2**
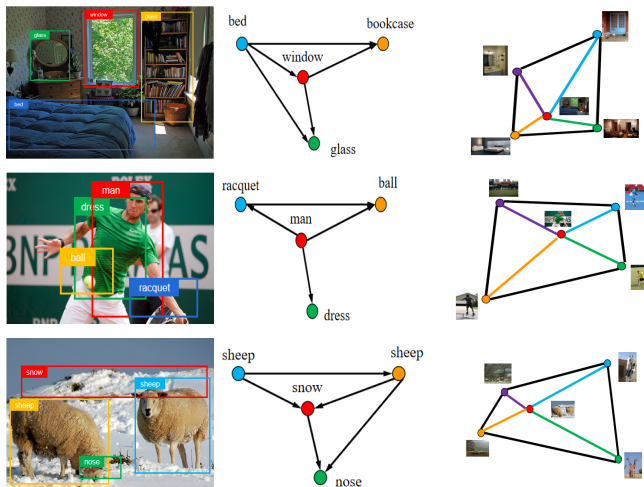
[Up-down] Bedroom scene with a bookcase and a window.

[AOA-net] A bed room with a bookshelf full of books.

[GCN-LSTM] A bed room with a large bookcase near a window.

[M^2 Transformer] A bed room with a neatly bed and a book shelf.

[Dual-GCN+Trans+CL] A room has a bed with *blue sheets and a large bookcase.*

[Up-down] A tennis player hits the tennis ball.

[AOA-net] A picture of a tennis player about to hit a ball.

[GCN-LSTM] A tennis player is hitting a tennis basll.

[M^2 Transformer] A tennis player is swinging at a tennis ball.

[Dual-GCN+Trans+CL] A man *holding a tennis racquet* in front of a tennis ball.

[Up-down] Two sheep standing in the snow.

[AOA-net] Two sheep standing next to each other in the snow.

[GCN-LSTM] Two sheep standing and eating in the snow.

[M^2 Transformer] Two sheep stand in the snow and has nose to the ground.

[Dual-GCN+Trans+CL] Two sheep standing in the snow *with one looking for food.*

Figure 4: Visualization results compared with state-of-the-art methods. Note that we only plot top-4 images for simplicity.

no image will appear in both MS COCO datasets and the Visual genome dataset at the same time.

Evaluation Metrics. We adopt 4 widely used metrics to evaluate the performance of our proposed model on the image captioning task, including BLEU-n [40], ROUGE [30], METEOR [7] and CIDEr [44]. The BLEU-n metric evaluates the similarity between the generated caption and the reference caption. The choice of parameter n refers to the reference caption in a different phrase. Therefore, a smaller number chosen as n means a better accuracy achieved by the generated caption; While a larger number of n stands for a better semantic similarity between the generated caption and the reference. BLEU-1 and BLEU-4 are the most commonly used indicators in terms of accuracy and semantic similarity. Except for uni-gram precision, METEOR has adopted a weighted harmonic mean to evaluate the performance. ROUGE uses the longest captions between the generated caption and reference caption to measure their similarity at the caption level. CIDEr is a special one measuring human consensus for caption generation.

4.2 Implementation Detail

Based on the Visual Genome dataset, we pretrain our object detector which is built upon the Faster-R-CNN [41] and enhanced by replacing its backbone with a resnet-101 [14]. For each input image, we firstly apply this detector to detect the object regions. Each region is represented by a 2048-dimensional vector $C = 2048$. Then we select the top $O = 36$ regions with the highest confidences to represent the image. In the transformer decoder part, we use a

1000 dimensional word embedding to represent each word instead of using one-hot vectors. And we add the sinusoidal positional encodings into the word embedding. The number of transformer encoder layers and decoder layers are both set to 6. Similar to [43], the dimensionality of the layer and the number of the muti-head attention are set to 512 and 8, respectively. The number of similar images K is set to 6. Toward the curriculum learning strategy, we calculate the average of BLEU-1 and BLEU-4 as the final score which turns out to be a more stable and realistic difficulty evaluation. We choose $M = 8$ as the number of the datasets.

4.3 Comparisons with State-of-the-Art

Compared Methods. We compare our proposed DGCN with six state-of-the-art methods including (1) two attention based methods, *i.e.*, Up-down [1] and AOA Net [21], (2) three graph based methods, *i.e.*, GCN-LSTM [55], GCN-LSTM+HIP [56] and SGAE [52], and (3) one transformer based method, M^2 transformer [5].

Quantitative Comparison. Table 2 shows the testing result of different models on the MS COCO dataset for image captioning task. Our proposed *Dual-GCN + Transformer + CL* model achieves the best results in all seven different evaluation metrics. Notably, it produces an 82.2 BLEU-1 score, which significantly outperforms other state-of-the-art models. This score is 2.0 higher than the traditional up-down method [1]. Compared with the latest AOA method [21], our model produces better BLEU-1 and CIDEr scores, which are 1.2 and 2.3 higher than AOA, respectively. We also observe that the

Table 3: Effectiveness comparison experiment on the MSCOCO dataset. Dual-GCN means the combination of GCN_{obj} and GCN_{img}. F_{obj} and F_{img} indicate the input features fed into GCN_{obj} and GCN_{img}, respectively. CL means Curriculum learning.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
F_{obj}+Transformer	71.1	55.9	45.3	28.0	25.2	52.8	114.3
GCN_{obj} + Transformer	80.5	64.7	49.2	37.6	27.9	57.4	125.8
F_{obj}+Transformer+CL	72.3	56.8	46.1	28.4	25.4	53.0	114.7
GCN_{obj} + Transformer +CL	81.6	66.8	51.6	39.3	28.3	58.4	**129.4**
F_{img} + Transformer + CL	80.2	64.1	49.0	37.4	27.4	57.0	124.0
GCN_{img} + Transformer + CL	82.0	66.9	52.2	**39.8**	29.6	59.4	129.0
$GCN_{obj}\&F_{img}$ + Transformer + CL	81.9	67.3	52.0	39.5	29.4	59.6	**129.4**
$GCN_{img}\&F_{obj}$ + Transformer + CL	82.0	67.3	52.3	39.7	**29.8**	59.6	129.0
Dual-GCN + Transformer	81.8	66.2	51.6	39.2	28.7	58.9	128.0
Dual-GCN + LSTM + CL	81.4	65.8	51.3	38.8	28.0	58.4	127.6
Dual-GCN + Transformer + CL	**82.2**	**67.6**	**52.4**	39.7	29.7	**59.7**	129.2

self-attention mechanism in our transformer encoding and decoding process helps to improve the performance as compared to the traditional LSTM decoding used by GCN-LSTM [55]. This is also justified by the fact that the score obtained by GCN-LSTM+HIP [56] is higher than that of the traditional LSTM based method [55]. However, the score of our model is higher than both GCN-LSTM [55] and GCN-LSTM+HIP [56]. Finally, different from M^2 transformer [5] which uses memory blocks, our proposed *Dual-GCN + Transformer + CL* model, which uses the Dual-GCN network, achieves the higher performance in terms of all the metrics. Obviously, these observations strongly demonstrate the efficacy of our proposed method.

Qualitative Analysis. Figure 4 shows the qualitative results compared with four state-of-the-art methods. We can clearly see that our model produces a caption that correctly describes the content of the input image. Especially, the captions generated by our *Dual-GCN + Transformer + CL* model are more detailed and authentic.

4.4 Ablation Study

We conduct an ablation study to verify the effectiveness of different components in our method. To this end, we introduce several variants of our full model as shown in Table 3: (i) *Transformer* means only using the transformer and the proposed region embeddings to generate image captions. (ii) F_{obj} means the object feature of the original image. (iii) F_{img} means the mean pooling result of the F_{obj}. (iv) GCN_{obj} means adopting the *GCN* to process the region embeddings before feeding them into the transformer. (v) GCN_{img} means adopting the *GCN* on F_{img} before feeding them into the transformer. (vi) *CL* means improving the previous model variant by using the Curriculum Learning strategy. (vii) $GCN_{obj}\&F_{img}$ and $GCN_{img}\&F_{obj}$ are two variants of our Dual-GCN in which both object-level and image-level visual features are used with only one GCN, either GCN_{obj} or GCN_{img}. (viii) *LSTM* means only using LSTM to replace the transformer to generate image captions. (ix) *Dual-GCN + Transformer + CL* is our full model and employs the proposed Dual-GCN module with transformer and curriculum learning to improve the quality of the final results further.

Effectiveness of Dual-GCN. The proposed Dual-GCN network aims at enhancing the latent embedding. To verify the effectiveness of our Dual-GCN in improving the performance of image captioning. We have conducted a comparative experiment on whether the GCN

network is used or not. The quantitative results are listed in Table 3. When we use $GCN_{obj}\&F_{img} + Transformer + CL$, it shows that the BLEU-1 and BLEU-2 score reaches 81.9 and 67.3 respectively. Morever, when we use $GCN_{img}\&F_{obj} + Transformer + CL$, the score of BLUE-1 and BLEU-2 reaches 82.0 and 67.3. But until we added Dual-GCN in our model, the result turned out to be better: the BLEU-1 and BLEU-2 scores reach 82.2 and 67.6 respectively, which is higher than the model that only adopts one GCN network. This shows that our Dual-GCN does have a certain improvement effect on the model.

Effectiveness of Transformer. The transformer has been used in image captioning only in recent years. To verify the difference between transformer and traditional LSTM, we added the contrast experiment between transformer and LSTM in our experiment. The quantitative results are listed in 3. When using *Dual-GCN+LSTM+CL*, the BLEU-1 score reaches 81.4 and the CIDEr reaches 127.6. However, when we replace the original LSTM with the transformer, the result changes a lot: The BLEU-1 score reaches 82.2 and the CIDEr score reaches 129.2, which is 0.8 and 1.6 higher than *Dual-GCN+LSTM+CL* respectively. At the same time, the other of the evaluation scores also have been improved. Therefore, transformer does have a better effect than LSTM in our model.

Effectiveness of Curriculum Learning. To verify the effectiveness of our proposed Curriculum Learning strategy, we have conducted a comparative experiment to train the transformer + GCN network either using curriculum learning or traditional training scheme. The experimental results are shown in Table 3. We can find that Curriculum Learning scheme introduces a specific improvement to our model. For example, without curriculum learning, the BLEU-1 score and BLEU-4 score only reached 81.4 and 38.8. When we adopt the curriculum learning to our *Dual-GCN+transformer* model, it studies from easy to complex and the BLEU-1 and BLEU-4 reaches 82.2 and 39.7. This shows the effectiveness of our curriculum learning training method.

To make readers better understand our proposed method, we also provide visualization results for comparison in Figure 5. It shows the ground truth caption and different generated captions of our model in ablation experiments. For simplicity, here we only plot top-4 images.
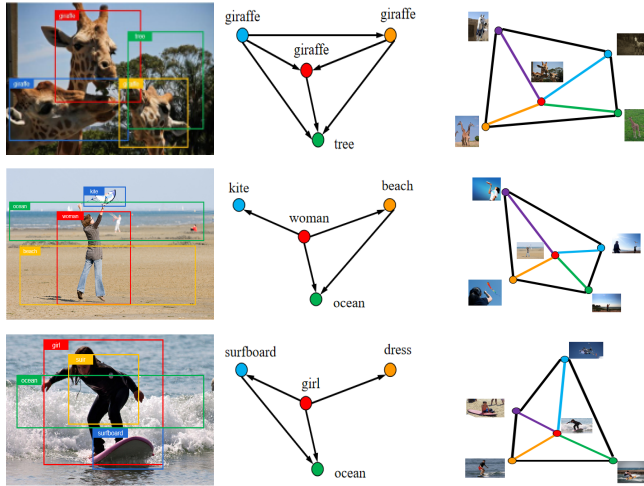
[GT] Three giraffes are eating and grooming each oher in the zoo.
[GCN_{img}&F_{obj}+Trans+CL] There are three giraffes are eating and standing.
[GCN_{obj}&F_{img}+Trans+CL] Three giraffes are eating near a tree.
[Dual-GCN+Trans] Three giraffes standing and eating near a tree.
[Dual-GCN+LSTM+CL] Three giraffes standing and eating next to each other.
[Dual-GCN+Trans+CL] Three giraffes standing and eating near trees.

[GT] A woman jumping up with a kite in her hands.
[GCN_{img}&F_{obj}+Trans+CL] A person playing a kite on beach.
[GCN_{obj}&F_{img}+Trans+CL] A woman holding a kite on beach.
[Dual-GCN+Trans] A woman jumping and playing a kite on beach.
[Dual-GCN+LSTM+CL] A person jumping and playing a kite on beach.
[Dual-GCN+Trans+CL] A person jumpling to hold a kite on beach.

[GT] A young girl tries to keep her balance on surfboard.
[GCN_{img}&F_{obj}+Trans+CL] A girl playing on a surfboard on beach.
[GCN_{obj}&F_{img}+Trans+CL] A girl playing on a surboard in the ocean.
[Dual-GCN+Trans] A young girl standing on a surboard in the ocean.
[Dual-GCN+LSTM+CL] A girl riding a surfboard in the ocean.
[Dual-GCN+Trans+CL] A young girl riding a surboard in the ocean.

Figure 5: Visualization results of ablation experiments. Note that we only plot top-4 images for simplicity.

4.5 Discussion

Different Quantities of K Similar Images. In similarity evaluation part, we need to choose how many similar images should be used to generate the caption. To determine the value of parameter K, we compare the performance by using different values of parameter K. Table 4 shows the quantitative result. We can see that taking $K = 6$ produces the best performance. This value is neither the smallest one nor the biggest one. This because smaller K affects the accurate information of a single image, while more images cover more image features. When K is set to 6, the score of BLEU-1 and BLEU-4 can reach the highest 82.2 and 39.7.

Table 4: The influence of different values of K.

	B-1	B-2	B-3	B-4	METEOR	ROUGE	CIDEr
K=3	80.2	64.8	49.4	37.2	27.4	57.0	124.7
K=4	80.7	65.0	50.2	37.4	27.9	57.4	124.9
K=5	81.6	66.2	51.3	38.6	29.0	59.0	126.8
K=6	**82.2**	**67.6**	**52.4**	**39.7**	**29.7**	**59.7**	**129.2**
K=7	82.0	67.1	52.2	39.4	29.4	59.5	128.9
K=8	81.8	66.7	51.7	38.9	29.2	59.2	127.6
K=9	82.0	67.0	51.9	39.1	29.4	**59.7**	128.4

Different Quantities of M Sub-datasets. To make sure the specific influence of different parameter M on the experimental results, we have tested our model using different integer values. Table 5 shows the quantitative result. The experimental results show the effect of increasing M on the performance of our model. In a specific range, the scores obtained by our model increase with the increase of M. But the scores go down when the value of M is bigger than a certain value. And our model obtains the best performance with $M = 8$. We take $M = 8$ in all other experiments.

Limitation and Failure Cases. To clarify, the measurement of similarity between different images largely depends on the $\mathcal{L}_2$ distance of the image feature vector. Therefore, when our algorithm wrongly selects images to guide the image captioning of the original image, it will have some negative effects. Figure 6 shows the visualization of negative examples.

Table 5: The influence of different values of M.

	B-1	B-2	B-3	B-4	METEOR	ROUGE	CIDEr
M=5	81.7	66.5	51.4	38.7	28.9	59.0	128.2
M=6	82.1	67.0	52.3	39.4	29.4	59.6	128.9
M=7	81.8	66.6	51.6	39.0	29.0	59.1	123.2
M=8	**82.2**	**67.6**	**52.4**	**39.7**	**29.7**	**59.7**	**129.2**
M=9	81.6	66.4	51.2	38.4	28.7	58.7	127.4

Input Image

[GT] *A bunch of red apples and yellow oranges in a closeup.*
[Output] *A red apple sitting on top of two apples.*

Figure 6: The result of similar image selection failure. When guiding the generation of apple, the model chooses the man in red clothes as the guide image with high similarity.

5 CONCLUSION

We propose a novel image captioning model in a combination of Dual-GCN and Transformer with Curriculum Learning as training strategy. The visual feature encoded by an object-level GCN and an image-level GCN is designed to incorporate both local and global visual encoding. The transformer decoder is able to understand the extracted visual features to generate reasonable caption results. With a cross-review mechanism to determine the difficulty of the dataset, we take curriculum learning as the training strategy to ensure that our proposed model is trained in a manner from easy to hard. Experiments conducted on the MS COCO dataset have strongly confirmed the efficacy of our proposed method.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 6077–6086.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. Association for Computing Machinery (ACM), 41–48.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 6298–6306.

[4] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 8299–8308.

[5] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 10578–10587.

[6] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting Visual Relationships with Deep Relational Networks. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 3298–3308.

[7] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (ACL)*. 376–380.

[8] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 2625–2634.

[10] Fei Fang, Fei Luo, Hongpan Zhang, Huajian Zhou, and Chunxia Xiao. 2020. A Comprehensive Pipeline for Complex Text-to-Image Synthesis. *Journal of Computer Science and Technology(JCST)* 35, 3 (2020), 522–537.

[11] Fei Fang, Miao Yi, Hui Feng, Shenghong Hu, and Chunxia Xiao. 2018. Narrative Collage of Image Collections by Scene Graph Recombination. *IEEE Transactions on Visualization and Computer Graphics(TVCG)* PP, 99 (2018), 2559–2572.

[12] Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043* (2017).

[13] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. 2018. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 11214. 139–154.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[15] Sen He, Wentong Liao, Hamed R Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. 2020. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, Vol. PP. 153–169.

[16] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*, Vol. 32.

[17] Tao Hu, Chao Liang, Geyong Min, Keqin Li, and Chunxia Xiao. 2020. Generating video animation from single still image in social media based on intelligent computing. *Journal of Visual Communication and Image Representation(JVCI)* 71 (2020), 102812.

[18] Tao Hu, Chengjiang Long, and Chunxia Xiao. 2021. CRD-CGAN: Category-Consistent and Relativistic Constraints for Diverse Text-to-Image Generation. (2021). arXiv:2107.13516

[19] Tao Hu, Chengjiang Long, and Chunxia Xiao. 2021. A Novel Visual Representation on Text Using Diverse Conditional GAN for Visual Recognition. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 3499–3512.

[20] Gang Hua, Chengjiang Long, Ming Yang, and Yan Gao. 2018. Collaborative Active Visual Recognition from Crowds: A Distributed Ensemble Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 40, 3 (2018), 582–594.

[21] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on Attention for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4633–4642.

[22] Ashraful Islam, Chengjiang Long, and Richard Radke. 2021. A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[23] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li Jia Li, and Li Fei-Fei. 2018. Mentor-net: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[24] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 4565–4574.

[25] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 39, 4 (2017), 664–676.

[26] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceeding of the International Conference on Learning Representations (ICLR)*. 5425–5434.

[27] 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. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)* 123, 1, SI (2017), 32–73.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* 60, 6, 84–90.

[29] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled Transformer for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 8927–8936.

[30] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. 74–81.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C.Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 8693. 740–755.

[32] Chengjiang Long, Arslan Basharat, , and Anthony Hoogs. 2019. A Coarse-to-fine Deep Convolutional Neural Network Framework for Frame Duplication Detection and Localization in Forged Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1–10.

[33] Chengjiang Long, Roddy Collins, Eran Swears, and Anthony Hoogs. 2019. Deep Neural Networks In Fully Connected CRF For Image Labeling With Social Network Metadata. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1607–1615.

[34] Chengjiang Long and Gang Hua. 2015. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2839–2847.

[35] Chengjiang Long and Gang Hua. 2017. Correlational Gaussian Processes for Cross-domain Visual Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 118–126.

[36] Chengjiang Long, Gang Hua, and Ashish Kapoor. 2016. A Joint Gaussian Process Model for Active Visual Recognition with Expertise Estimation in Crowdsourcing. *International Journal of Computer Vision (IJCV)* 116, 2 (2016), 136–160.

[37] Chengjiang Long, Eric Smith, Arslan Basharat, and Anthony Hoogs. 2017. A C3D-based Convolutional Neural Network for Frame Dropping Detection in a Single Video Shot. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1898–1906.

[38] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 3242–3250.

[39] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The More You Know: Using Knowledge Graphs for Image Classification. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 20–28.

[40] K Papineni, S Roukos, T Ward, and WJ Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 311–318.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 39, 6 (2017), 1137–1149.

[42] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. 2021. SGCN: Sparse Graph Convolution for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8994–9003.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*, Vol. 30.

[44] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 4566–4575.

[45] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 3156–3164.

[46] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 39, 4

(2017), 652–663.

[47] Hua Wang, Heng Huang, and Chris Ding. 2011. Image Annotation Using Bi-Relational Graph of Images and Semantic Labels. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 793–800.

[48] Jun Xiao, Lin Li, Dejing Xu, Chengjiang Long, Jian Shao, Shifeng Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Explore Video Clip Order with Self-Supervised and Curriculum Learning for Video Applications. *IEEE Transactions on Multimedia (T-MM)* (2020).

[49] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum Learning for Natural Language Understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 6095–6104.

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Y. Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Computer Science* (2015), 2048–2057.

[51] Wenju Xu, Shawn Keshmiri, and Guanghui Wang. 2019. Adversarially approximated autoencoder for image generation and manipulation. *IEEE Transactions on Multimedia (T-MM)* (2019), 2387–2396.

[52] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 10677–10686.

[53] Bangpeng Yao and Li Fei-Fei. 2010. Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 9–16.

[54] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 5263–5271.

[55] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 11218. 711–727.

[56] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy Parsing for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2621–2629.

[57] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting Image Captioning with Attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4904–4912.

[58] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 4651–4659.

[59] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Haiyin Piao, Haiyang Mei, Xin Yang, and Baocai Yin Yin. 2021. A Two-Stage Attentive Network for Single Image Super Resolution. *IEEE Transactions on Circuits and Systems for Video Technology* (2021).

[60] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Xin Yang, Haiyang Mei, and Baocai Yin. 2020. Multi-Context And Enhanced Reconstruction Network For Single Image Super Resolution. In *Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

[61] Yaohui Zhu and Shuqiang Jiang. 2018. Deep Structured Learning for Visual Relationship Detection. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*. 7623–7630.