# What Happens in Crowd Scenes: A New Dataset about Crowd Scenes for Image Captioning

Lanxiao Wang, Hongliang Li, *Senior Member, IEEE*, Wenzhe Hu, Xiaoliang Zhang, Heqian Qiu, Fanman Meng, *Member, IEEE*, Qingbo Wu, *Member, IEEE*,

***Abstract*—Making machines endowed with eyes and brains to effectively understand and analyze crowd scenes is of paramount importance for building a smart city to serve people. This is of far-reaching significance for the guidance of dense crowds and accident prevention, such as crowding and stampedes. As a typical multimodal scene understanding task, image captioning has always attracted widespread attention. However, crowd scene understanding captioning is rarely studied due to the unobtainability of related datasets. Therefore, it is difficult to know what happens in crowd scenes. In order to fill this research gap, we propose a crowd scenes caption dataset named CrowdCaption which has the advantages of crowd-topic scenes, comprehensive and complex caption descriptions, typical relationships and detailed grounding annotations. The complexity and diversity of the descriptions and the specificity of the crowd scenes make this dataset extremely challenging to most current methods. Thus, we propose a Multi-hierarchical Attribute Guided Crowd Caption Network (MAGC) based on crowd objects, actions, and status (such as position, dress, posture, etc.) aiming to generate crowd-specific detailed descriptions. We conduct extensive experiments on our CrowdCaption dataset, and our proposed method reaches the state-of-the-art (SoTA) performance. We hope the CrowdCaption dataset can assist future studies related to crowd scenes in the multimodal domain. Our dataset and the code of the benchmark method will be released soon.**

***Index Terms*—CrowdCaption, Image Captioning, Crowd Scenes, Multimodal Understanding.**

## I. Introduction

With the development of global science and technology, how to use large amounts of digital visual information about crowd scenes to serve people is of far-reaching significance. However, it is laborious to process and analyze the continuous flow of image and video information about crowd scenes. Thus, strong technical support is required to complete the transition from an informational city to a digital one and eventually a smart one. Under this circumstance, it is critical to provide machine eyes and brains to effectively understand and analyze crowd scenes, which can be used for the guidance and management of dense crowds and the prevention of accidents such as crowding and stampedes. This can also assist with smart cities, intelligent transportation, public security, and other application scenarios related to crowd scenes.

In recent years, an increasing number of researchers have focused on the study of crowd-related tasks, including crowd detection [6]–[8], crowd counting [9], [10], crowded scenes pose estimation [11], [12], and crowd tracking [13], [14]. Some researchers also proposed corresponding crowd-topic datasets, such as Crowd Dyadic Dialogues Dataset [15], Motion-Centered Figure Skating Dataset [16], CrowdPose [11], CrowdHuman [17], and DroneCrowd [18]. However, no one has proposed a dataset for crowd scenes descriptions.

Fig.1 shows some previous image datasets and our CrowdCaption dataset. As typical common datasets, COCO [1] and Flickr30k [2] are comprehensive and include images of various types and topics. As shown in (a) and (b), the five captions marked are very similar and only focus on the same significant object. However, for complex scenes, it is difficult to ensure that the objects that all people hope to describe are the same. For example, the object that some people care about is inconspicuous in the image, which is not contained in the caption annotations. In addition, the marked sentence structure is relatively singular. Fig.1 (c) shows an example of VisualGenome [3], which contains lots of bounding boxes with simple sentences or phrases. These annotations are more concerned with single objects such as mountains, pants, sky, and men. Fig.1 (d) and (e) show two caption datasets with a single topic: (d) an emotional description of the image, and (e) is aimed at the text topic image.

We observe that there are four limitations in existing image captioning datasets for the study of crowd scenes understanding: 1) only a small number of images are relevant to crowd scenes, 2) the diversity of views toward the same image is neglected, 3) the single-sentence structure and the descriptions with a single object cause it to be unable to describe complex crowd scenes, and 4) images often contain a prominent object and a clean background, which is different from real life. Thus, in order to improve the understanding and research of crowd scenes, it is urgent to build a dataset of image descriptions of crowd scenes. To fill in this gap, we propose a crowd scenes image captioning dataset named CrowdCaption.

Different from the above datasets, our CrowdCaption dataset is aimed at the crowd scenes and has practical social significance in the context of the growing population and the increasingly rich visual information. CrowdCaption can be used for crowd-specific scene understanding assisting the management and service of smart cities, intelligent transportation, public security, and other application scenes related to dense crowds. All images are collected based on twenty crowded places and gathering activities in real life. Not only the crowd topic of images but also the comprehensive and diverse interrelated sentences and large number of attribute features

The authors are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: lanxiao.wang@std.uestc.edu.cn; hlli@uestc.edu.cn; wenzhe-hu@std.uestc.edu.cn; xlzhang@std.uestc.edu.cn; hqqiu@std.uestc.edu.cn; fmmeng@uestc.edu.cn; qbwu@uestc.edu.cn).

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3192729



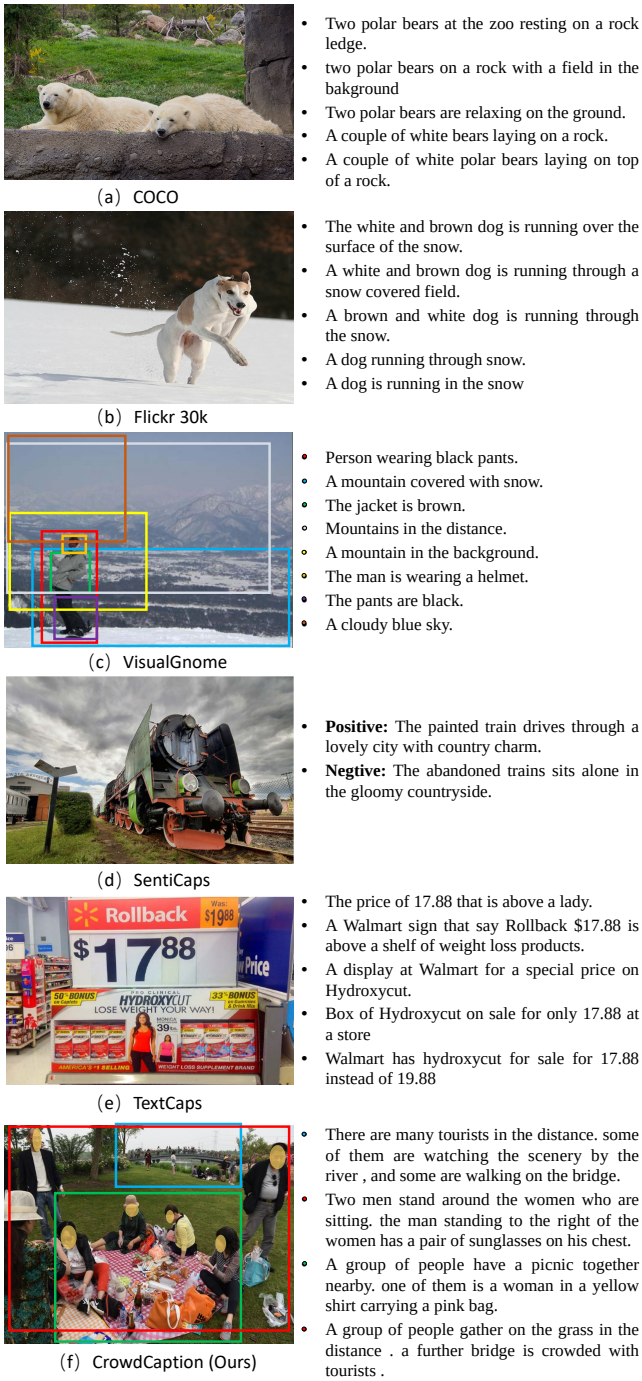



Fig. 1: Examples in CrowdCaption and previous datasets regarding image captioning, including COCO [1], Flickr30k [2], VisualGenome [3], TextCaps [4] and SentiCap [5]. CrowdCaption focuses on describing the actions and states of crowd scenes, focusing on individual people and relationships within crowds.

(crowd topic, action, position, dress, posture, surroundings, group characteristics, etc.) add value to our dataset.

In addition, as shown in Fig.1 (f), some annotations focus on tourists in the distance, some annotations focus on standing men, and some annotations focus on crowds having a picnic. Our captions in the same image are no longer limited to describing the same meaning and object, which follows the idea that different people have different focuses. Considering the potential application value of the dataset, we also annotate prominent crowd regions in images and human objects in the regions to provide more auxiliary information, which also has practical research significance. Researchers can choose different auxiliary information according to the actual research situation.

Compared with current caption-related datasets, our dataset has the following remarkable advantages: 1) all images in our dataset are pertinent to crowd scenes, containing rich real-life scenes, 2) the annotations of images are all comprehensive and diverse interrelated sentences, covering plenty of features including action, position, dress, posture, surroundings, and group characteristics, 3) our description focuses on different individuals or groups in complex images, and is no longer limited to describing the same content, 4) we label the crowd regions and objects corresponding to the descriptions, which can provide latent contextual information in the process of generating descriptions, and this information can bring more possibilities for future research, and 5) we carefully examined the quality of the selection and annotation of images to ensure that all images were in line with the real situation with practical research value, and all annotations were correct and reliable.

In order to better describe the crowd scenes, we also propose a benchmark method called the Multi-hierarchical Attribute Guided Crowd Caption Network to achieve fine-mapping from vision to language and generate crowd-specific detailed descriptions. Specifically, we firstly use a multi-hierarchical crowd aware module to extract different hierarchical crowd attribute features based on the crowd attribute loss. Then, we design a dynamic fusion module to select and integrate the multi-hierarchical visual information, on which the current decoding stage focuses. Finally, according to the high-level visual features obtained by the dynamic fusion module, we design a decoder to leverage these features to predict corresponding words step by step.

In summary, the contribution of our work can be listed as follows:

- To the best of our knowledge, we construct the first crowd scenes image captioning dataset in multimodal task, named CrowdCaption. It is used to understand crowd scenes of an entire image and generate captions related to visual crowd scenes from multiple special perspectives. Exhaustive analysis and comprehensive comparison show the superiority of our dataset in image captioning related to crowd scenes understanding.
- We propose a benchmark method for CrowdCaption named the Multi-hierarchical Attribute Guided Crowd Caption (MAGC) Network. This achieves fine mapping from vision to language based on crowd objects, actions, and other statuses and generates crowd-specific detailed descriptions.
- We conduct extensive experiments with our benchmark and state-of-the-art methods on CrowdCaption. These experiments prove the effectiveness of our method.

## II. Related Work

### A. Datasets

In order to explore the relationship between vision and language, a large number of datasets have been proposed for image captioning task. COCO [1] and Flickr30k [2], as the most common datasets in the image captioning task, contain diverse images, and each image contains five captions. However, nearly all five annotations for each image have the same meaning, even if the image is fairly complex and contains much information. Plummer et al. [19] proposed Flickr30k Entities based on Flickr30k building the link between entities words and bounding boxes. Krishna et al. [3] proposed Visual Genome, which contains many bounding boxes on images and describes its content with a phrase or simple sentence. This is typically used for dense caption task. In order to generate a coherent natural language description for images, Krause et al. [20] collected an image paragraph dataset named the Stanford image-paragraph. This dataset aims to generate a detailed and unified story. The average length of descriptions in this dataset is 67.5. Due to the long paragraphs and lack of auxiliary information such as detection annotations, this dataset is still difficult to use in research.

With the gradual development of image captioning task, researchers have focused on controllable descriptions and more targeted single-topic descriptions, which pays more attention to human thoughts and needs. Mathews et al. [5] hoped to generate emotion controlled captions and proposed the SentiCap dataset, which contains 3 positive and 3 negative captions for each image. In 2020, many targeted caption datasets appeared. Sidorov et al. [4] thought that text is ubiquitous in the human environment, and generating captions based on the text in the image is crucial. Thus, they proposed the TextCaps dataset to fill the gap in the research field. Yang et al. [21] built the Fashion Captioning Dataset [21] to provide services for customers' online shopping and enhance the increase in online sales.

Different from these datasets, our CrowdCaption dataset is aimed at the crowd scenes. Zheng et al. [22] thought that it is difficult to guarantee that an object people care about is contained in descriptions. Our captions are no longer limited to describing the same meaning for the same image, which is more in line that different people have different focuses. Our captions consist of comprehensive and diverse interrelated sentences, and annotations contain a large amount of attribute information about crowds and prominent crowd regions and objects.

### B. Methods

Image captioning has aroused wide concern in the computer vision area since Vinyals et al. [23] used the encoder-decoder structure to generate word sequences based on RNNs. To achieve the interaction between modalities, later methods focused on utilizing the attention mechanism. Some researchers [24], [25] made use of an attention mechanism to select the most relevant image area in each decoding stage of LSTM. Anderson et al. [26] proposed a bidirectional attention idea. In 2018, the bottom-up attention model was used to extract regions of interest in images to obtain object features, whereas the top-down attention model was employed to learn weights corresponding to the features to accomplish an in-depth understanding of visual images. In the same year, Li et al. [23] proposed global-local attention to select object-level integration based on image-level features. Yao et al. [27] paid their attention on the connections between objects and built graphs based on visual relationships between spatial and semantic connections to generate captions. In 2019, AoANet [28] extended the conventional attention mechanism by taking into account the relevance of attention and query results. Yang et al. [29] tried to simultaneously optimize image captioning and text-to-image synthesis based on the idea of multitasking. In order to fully use the vertical depth advantages of the network, a deep hierarchical encoder-decoder network [30] was proposed, which can fuse high-level features of vision and language. Moreover, many methods [31], [32] construct triples based on scene graphs to fully mine structured information to assist in the decoding of language.

In recent years, with the further development of attention mechanisms, researchers are no longer focused only on simple global or local attention, but have studied more complex and deep attention mechanisms [33]–[35]. For example, Pan et al. [33] designed an X-Linear attention block that adopted bilinear fusion to mine second-order or even higher-order feature interaction between modalities to enhance cross-modal content understanding. In addition, inspired by the field of natural language processing, some researchers [36], [37] began to pay attention to sentence structure and Part-of-Speech, who based on Part-of-Speech to build and extract more effective visual and language features. Some researchers believe that the accumulation of errors will lead to a failed description [38], [39]. CaptionNet [38] proposed an improved LSTM, which only allows the image features of interest to pass through, thereby reducing the impact of preorder words. Guo et al. [39] built a ruminant caption framework to mimic the human polishing process to guide the prediction of words. In order to reduce the problem that the network only generates frequent words caused by the long tail effect, Guo proposed a global-local discriminative objective to better distinguish the corresponding image from all images and emphasize words with less frequency.

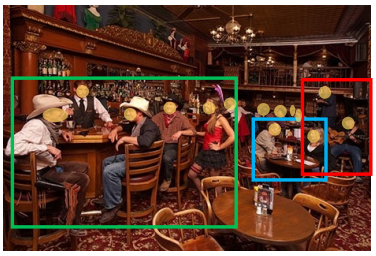
Although the abovementioned methods achieved advanced performance in previous caption datasets, they are insufficient for the challenges of complex real-world crowd scenes. To address this problem, our method pays more attention to crowd characteristics, and extracts multi-hierarchical attribute features, which contains more crowd-specific detailed visual information. Then, we further use multi-hierarchical attribute features to guide the decoding process of sentences.

## III. CrowdCaption dataset

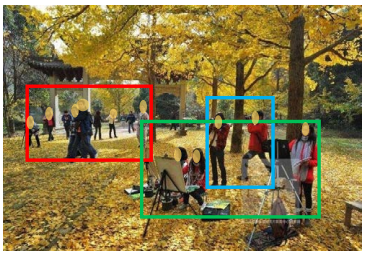
CrowdCaption aims to be a multimodal caption dataset with the theme of crowd scenes. Fig.2 shows examples in different crowd scenes. Compared with existing datasets, CrowdCaption has complex caption descriptions of crowd scenes. This descriptions contain two unrelated subsentences with typical

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3192729





- a woman in black coat is sitting at a table. a bald man sits beside her.
- two men are playing the instruments. the man on the chair is playing the guitar.
- three men with hats are sitting on the chair. a barman and a lady with a black skirt are standing beside them.

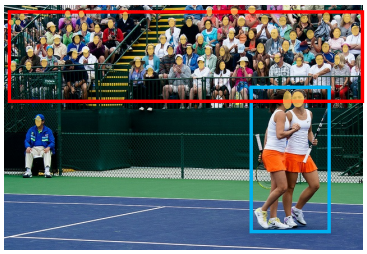


- some people are holding cameras. they are taking photos of the beautiful scenery.
- there are a few people walking in the left. they are enjoying the beautiful scenery.
- some people are drawing on the boards. the far right one sitting under tree wears glasses.

- there are many tourists in the distance. some of them are watching the scenery by the river , and some are walking on the bridge.
- two men stand around the women who are sitting. the man standing to the right of the women has a pair of sunglasses on his chest.
- a group of people have a picnic together nearby. one of them is a woman in a yellow shirt carrying a pink bag.

- there are a few women standing in the room. they are watching their children making handwork.
- there are a few children sitting in the chairs. they are making handwork.

- many elementary school students are standing on the left. a man in a blue jacket is directing them to cross the road.
- many children walk on the right side of the road. the little girl on the left is carrying a blue schoolbag.



- there are two female players on the court. they wear white vests and orange skirts.
- people are crowded in the stadium. they are watching a tennis match.

Fig. 2: Examples of typical crowd scenes in CrowdCaption. Different colors represent tendency region and objects of caption on the image.

crowd behaviors and relationships. In addition, we mark the region and objects that each sentence focuses on in the image, which is of great research value in crowd scenes research. We hope this dataset can promote AI research and make machines more intelligent to understand crowd scenes. In this section, we introduce the data collection, data annotation, quality control, and statistical analysis of this approach.

### A. Image Collection

We collect a total of 11,161 images of crowd scenes. First, we build an annotation team containing 24 human subjects who have 2 more years of deep learning research experience. After all human subjects learned the annotations of some existing datasets together, we conducted small-scale image collection and annotation to ensure the quality and uniformity of the dataset. After this, we searched COCO dataset [1] and selected 6,204 images for CrowdCaption. Considering the crowded places and gathering activities in real life, we further set twenty common crowd scenes as keywords, as shown in Table-I. Following some common datasets [1], [40], we acquired images from image searches on Google and Bing based on these keywords. Considering the balanced distribution of the image scene category, we collected 300 images for each keyword. We inspected all images and eliminated those which do not contain crowds or have no practical use. We also ensured that the scenes in the images are common in human life. Thus, all data have value for research. Finally, we gathered 4,957 images of crowd scenes.

TABLE I: Keywords in the processing of collecting crowd scenes images.

| Keywords | | | |
|---|---|---|---|
| sports | supermarket | street | hospital |
| station | seabeach | picnic | office |
| shopping | pub | meeting | sea/boat |
| park | tourist spot | outdoors | party |
| club | restaurant | classroom | carnie |

### B. Annotation Collection

All annotations in our dataset are created by human subjects. We choose the labeling tool LabelImg[1] to label and align

TABLE II: Comparison of different datasets about image captioning.

| Dataset | Images | Captions per image | Persons | Persons per image | Avg. Captions Length | Theme |
|---|---|---|---|---|---|---|
| COCO [1] | 123,287 | 5 | 31,783 | 1.46 | 10 | no single theme |
| Flickr30k [2] | 31,783 | 5 | 199,279 | 6.27 | 12 | no single theme |
| VisualGenome [3] | 108,007 | 1 per single object | 125,368 | 1.16 | 5 | no single theme |
| TextCaps [4] | 28,408 | 5.1 | - | - | 12 | captions about text in image |
| SentiCap [5] | 3,171 | 3 positive and 3 negative | - | - | 12 | captions with emotion controlled |
| CrowdCaption (Ours) | 11,161 | 4.4 | 186,500 | 16.71 | 20 | captions about crowds |

captions and the location annotation of objects and regions in the image. Human subjects need to follow three steps to accomplish the annotation work: language annotation, region annotation, and object annotation. First, human subjects need to create descriptions for different regions. Ensuring the diversity of language, we ask human subjects to create two captions for each region. For overly simple regions, labeling 1 caption is also allowed. Second, while labeling captions, human subjects need to label the prominent crowd regions in the image. Each region is required to contain at least 2 relevant people or a single salient person. Due to differences in the density of crowds in images, the number of regions for each image is also different. Finally, human subjects need to annotate the human objects in the image. Considering the labeling time, for a large number of gathering people such as spectators in stands, subjects are allowed to label the whole region containing these people rather than each individual.

In order to ensure the effectiveness and quality descriptions for crowd scenes, we set some requests as follows: 1) Subjects need to make diversified descriptions for different attention angles. 2) The caption needs to include rich information about human beings, such as attributes, behaviors, and status. 3) Some relationships, such as lovers and friends, are ambiguous in visual information. In order to avoid this ambiguity, subjects are requested to describe only apparent and typical relationships. 4) All captions need to avoid any ethical issues including pornography, violence, and gender and racial discrimination.

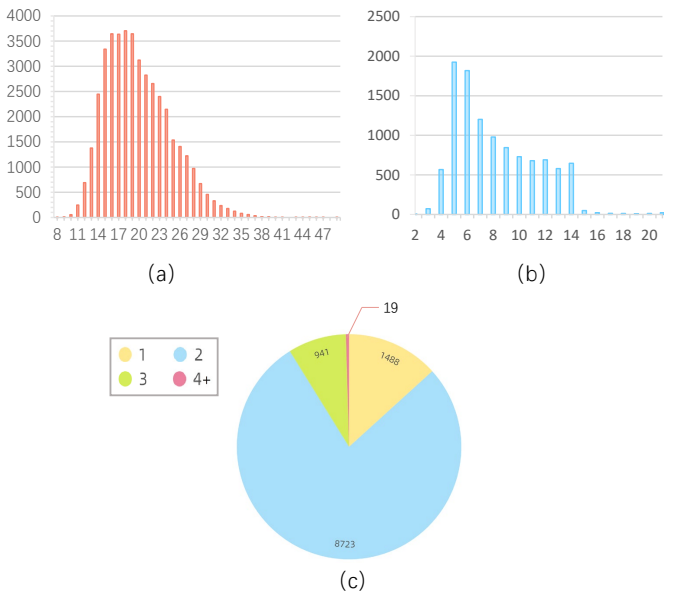



Fig. 3: Illustration of statistics about CrowdCaption: (a) distribution of length of sentence, (b) distribution of number of human object annotations per image, and (c) number of salient region annotations in image.

## C. Quality Control

We take a series of operations to ensure the quality of CrowdCaption. Following the principle of cross-checking, we invited 24 examiners and divided them into 4 groups to conduct quality control. First, we asked them remove all data that may be related to identity information, sexism or racism to ensure that all images come from public scenes to avoid personal privacy concerns. We asked them to modify the error of annotations to avoid a nondetailed description, location inaccuracy, label matching error, etc. All errors will be corrected. Moreover, we use a Python LanguageTool language-check as syntactic objective evaluations, which ensures that our annotations comply with basic syntactic rules.

For ethical considerations, CrowdCaption is created with careful attention to ethical considerations and only used for research purposes. We remove all images and annotations that may be related to pornography and violence during quality checks. Our dataset does not contain identity information or gender and racial discrimination in any annotations. We prioritize all public features and annotations. Researchers need to sign the CrowdCaption Terms of Use as restrictions to download images which can protect privacy protection, and we also allow people to contact us to delete specific images. We will also mask all face details in our paper.

[1]https://github.com/tzutalin/labelImg

## D. Dataset Statistic

Our CrowdCaption dataset contains 11,161 images with 43,306 captions, 21,794 regions and 95,820 objects. Each region has an average of 2 captions. The entire dataset is divided into training, validation, and testing sets of 7,161, 1,000, and 3,000 images, respectively.

Table-II shows a statistical comparison between our dataset CrowdCaption and existing image captioning datasets including COCO [1], Flickr30k [2], VisualGenome [3], TextCaps [4], and SentiCap [5]. Early image captioning datasets such as COCO, Flickr30k, and VisualGenome are comprehensive and focus on multiple categories. In recent years, an increasing number of special datasets have appeared such as TextCaps and SentiCap, which have a single theme or characteristic. Compared with these datasets, our CrowdCaption is a typical crowd-oriented dataset which only focuses on crowd scenes.


(a) all words (b) object-level words

(c) action-level words (d) status-level words

Fig. 4: Word cloud for our CrowdCaption dataset. Size of the word represents its frequency, including all words, words about object, words about object's action and object's status in (a), (b), (c), and (d). We remove all stop words and words appearing fewer than two times.

As shown in Table-II, we count the number of people in each image. Our datasets have an average of 16.71 people per image, which is almost an order of magnitude denser than any existing image captioning dataset. Therefore, existing datasets are difficult to apply to crowd scenes.

Moreover, we count the lengths of the captions in Fig.3 (a). It can be seen that our captions are longer and more comprehensive and diverse. This is because each of our annotations contains two interrelated sentences, and this kind of interrelated sentence contains more information, which is more in line with human description habits. As shown in Fig.3 (b), most of the images contain 5-14 human object annotations in the saliency regions. Fig.3 (c) shows the number of regions in each image, and the distribution illustrates the complexity of the image. Most images contain 2 regions, while only a few extremely dense and complex images contain 4 regions. These further analyses fully demonstrate that images in dense crowd scenes are very complex and challenging.

In order to illustrate the crowd specificity of our dataset more vividly, we perform word-level analysis on descriptions in the dataset. We first analyze all annotations using Stanford NLP tools, which is commonly used in NLP, to obtain part-of-speech (POS) annotations for each word. Then, after removing all stop words and words appearing fewer than two times, all words are divided into the object-level, action-level and status-level according to POS categories. Object-level mainly include noun categories, action-level mainly include verb categories, and status-level mainly include adjective categories. Finally, we build a word cloud based on the word frequency for different levels, which are shown in Fig.4. For (a), the whole words are analyzed. For (b), (c), and (d), the words show typical categorical properties. By observing different categories of words such as "man", "woman", "t-shirt" and "coat" in (b); "standing", "wearing" and "holding" in (c); and "left", "behind" and "white" in (d); it can be seen that our dataset has a detailed description covering the action, clothing, location, relationships, etc., details of the crowd, which is typically crowd-specific.

## IV. Method

Considering the attributes of the crowd scenes, we propose a Multi-hierarchical Attribute Guided Crowd Caption Network. First, a multi-hierarchical crowd aware module is proposed to extract the crowd attribute features of the corresponding hierarchic. We further build a dynamic fusion module as a part of the encoder. Finally, we use the high-level crowd attribute features obtained by the dynamic fusion module to guide the decoder step by step. Figure-5 shows the framework of the benchmark.

In this section, we introduce the crowd feature encoder, including the multi-hierarchical crowd aware module and high-level feature fusion module, which aim to enhance the crowd target pertinence of visual information. Then, the encoder can capture both global visual features and high-level crowd attribute features. Finally, the crowd guided decoder is used to perform multimodal inference and output descriptions.

### A. Notations

We define the sentence as $S_{1:T} = \{w_1, ..., w_T\}$, and $S_{1:T}$ contains T words, where $w_t$ is the t-th word. The visual features of image $I$ are extracted from pretrained Faster R-CNN [41], [42] as a set of region features $F_r = \{f_1^r, ..., f_N^r\}$ and position features $F_p = \{f_1^p, ..., f_N^p\}$. $N$ is the number of regions in the image. We utilize HRNet [43] to capture crowd posture features $F_c = \{f_1^c, ..., f_M^c\}$. $M$ is the number of joints in human pose estimation. $C_n$ is the label of the $n$-th attribute, including the object attribute, action attribute, and status attribute.

### B. Visual Feature Embedding

In order to capture the visual region information, we embed the detection feature $F_r$ into a fixed feature space and obtain $V_r$ as the main visual information.

$$\begin{aligned} MLP(*) &= LayerNorm(ReLU(FC(*))) \\ V_r &= MLP(F_r) \end{aligned} \tag{1}$$

where $F_r \in \mathbb{R}^{N\times 2048}$, and $V_r \in \mathbb{R}^{N\times D_{emb}}$.

Similarly, we obtain the position of the detection box information and embed it into the same feature space.

$$V_p = MLP(F_p) \tag{2}$$

where $F_p \in \mathbb{R}^{N\times 5}$, and $V_p \in \mathbb{R}^{N\times D_{emb}}$.

However, for the topic of crowd scenes, the embedding features $V_r$ and $V_p$ may not provide enough visual information related to the crowd. In order to further enhance the pertinence of regional features to the crowd, we add a human pose feature $F_c$. Human pose features are powerful supplements to the existing multi-target detection features, so we combine them to obtain richer crowd visual information.

We first flatten crowd posture features $F_c$. In order to extract deeper semantic information, we adopt a three-layer

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3192729



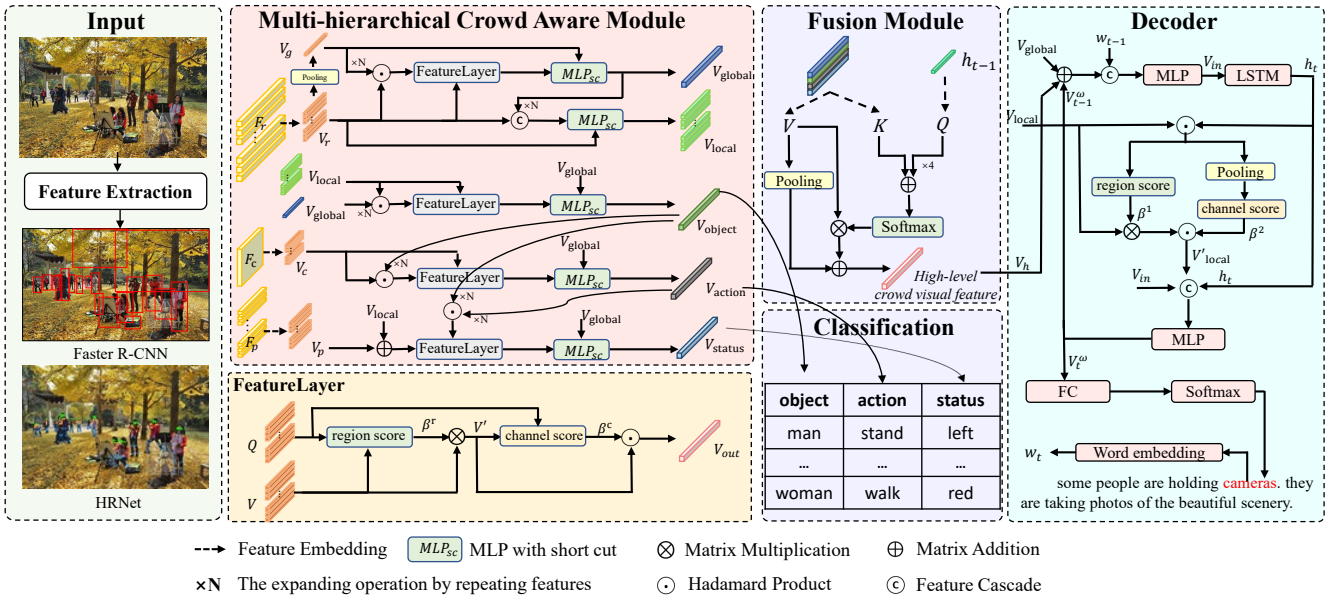



Fig. 5: The framework of benchmark method Multi-hierarchical Attribute Guided Crowd Caption Network on CrowdCaption dataset.

MLP structure to further exploit the complex human visual features $V_c$ in the image while increasing the nonlinearity of the features.

$$V_c = MLP_3(flatten(F_c)) \tag{3}$$

where $MLP_3$ represents the three-layer MLP structure, $F_c \in \mathbb{R}^{M\times 128\times 208}$, and $V_c \in \mathbb{R}^{M\times D_{emb}}$.

### C. Multi-hierarchical Crowd Aware Module

For crowd scenes image captioning, the most important function of the feature encoder is to enhance the object pertinence and regional pertinence of visual features. Therefore, it is particularly crucial to capture the key information of the concerned crowd scenes and the global information of the image.

In order to provide the decoder with richer visual features containing different attribute trends, we build a Multi-hierarchical Crowd Aware Module, including general-purpose feature extraction layer, object attribute feature extraction layer, action attribute feature extraction layer and status attribute feature extraction layer.

In detail, we use the idea of the spatial attention mechanism and channel attention mechanism to build an extraction layer to extract hierarchical visual features. We first define query as $Q$, value as $V$ and the extracting process as $FeatureLayer(Q, V)$.

$$V_{out} = FeatureLayer(Q, V) \tag{4}$$

where $Q \in \mathbb{R}^{K\times D}$, $V \in \mathbb{R}^{K\times D}$, and $V_{out} \in \mathbb{R}^{1\times D}$.

We use two embedding layers to obtain embedding features of $Q$ and $V$ in the same feature space, and then learn the regional score $\beta^r$ based on the features after $tanh$ processing.

$$\begin{aligned} \beta^r &= softmax(W_3 tanh(W_1 Q + W_2 V)) \\ V' &= \beta^r V \end{aligned} \tag{5}$$

where $W_*$ are trainable parameters, $\beta^r \in \mathbb{R}^{1\times K}$, and $V' \in \mathbb{R}^{1\times D}$.

Then, based on the region-weighted visual feature $V'$, we further extract channel-level visual score $\beta^c$ to enhance the expression ability of visual feature.

$$\begin{aligned} \beta^c &= softmax(W_6 tanh(W_4 AvgPool(Q) + W_5 V')) \\ V_{out} &= \beta^c \odot V' \end{aligned} \tag{6}$$

where $W_*$ are trainable parameters, $\odot$ represents the Hadamard Product, $\beta^c \in \mathbb{R}^{1\times D}$, and $V_{out} \in \mathbb{R}^{1\times D}$.

For the general-type feature extraction layer, we first obtain a low-level global features $V_g$ via average pooling.

$$V_g = AvgPool(V_r) \tag{7}$$

where $V_r \in \mathbb{R}^{N\times D_{emb}}$, and $V_g \in \mathbb{R}^{1\times D_{emb}}$.

We further use low-level global and regional features to extract more advanced and general global visual features $V_{global}$.

$$\begin{aligned} V_{query} &= [V_g]_N \odot V_r \\ V'_g &= FeatureLayer(V_{query}, V_r) \\ V_{global} &= MLP_{sc}(V'_g, V_g) \end{aligned} \tag{8}$$

where $MLP_{sc}(*_1, *_2) = LayerNorm(ReLU(FC(*_1) + *_2))$, $[*]_N$ indicates the expanding operation by repeating features, $V_{query} \in \mathbb{R}^{N\times D_{emb}}$, $V'_g \in \mathbb{R}^{1\times D_{emb}}$, and $V_{global} \in \mathbb{R}^{1\times D_{emb}}$.

Based on $V_{global}$, we supplement global visual information to obtain more high-level and general local visual information $V_{local}$. We concatenate these features and meanwhile employ a fully connected layer for cross-dimensional information

interaction. In addition, we use the shortcut structure for feature reuse while avoiding gradient loss.

$$\begin{aligned} V_r' &= [[V_{global}]_N, V_r]_c \\ V_{local} &= MLP_{sc}(V_r', V_r) \end{aligned} \tag{9}$$

where $[*,*]_c$ represents the feature cascading operation on the channel dimension, $V_r' \in \mathbb{R}^{N\times 2D_{emb}}$, and $V_{local} \in \mathbb{R}^{N\times D_{emb}}$.

For multi-hierarchical crowd features, we first embed global visual features into the object feature space.

$$\begin{aligned} V_{object\_query} &= [V_{global}]_N \odot V_{local} \\ V_{object\_att} &= FeatureLayer(V_{object\_query}, V_{local}) \\ V_{object} &= MLP_{sc}(V_{object\_att}, V_{global}) \end{aligned} \tag{10}$$

where $V_{object\_query} \in \mathbb{R}^{N\times D_{emb}}$, $V_{object\_att} \in \mathbb{R}^{1\times D_{emb}}$, and $V_{object} \in \mathbb{R}^{1\times D_{emb}}$.

Since behaviors and objects in the generated descriptions are interconnected, and human pose features contain rich crowd behavior information, we use object visual features and human pose features to build the behavior feature space.

$$\begin{aligned} V_{action\_query} &= [V_{object}]_N \odot V_c \\ V_{action\_att} &= FeatureLayer(V_{action\_query}, V_c) \\ V_{action} &= MLP_{sc}(V_{action\_att}, V_{global}) \end{aligned} \tag{11}$$

where $V_{action\_query} \in \mathbb{R}^{N\times D_{emb}}$, $V_{action\_att} \in \mathbb{R}^{1\times D_{emb}}$, and $V_{action} \in \mathbb{R}^{1\times D_{emb}}$.

Furthermore, we look for other status features spaces which are related to objects and behaviors such as position, color, and number.

$$\begin{aligned} V_{status\_query} &= [V_{object}]_N \odot [V_{action}]_N \\ V_{status\_att} &= FeatureLayer(V_{status\_query}, V_{local} + V_p) \\ V_{status} &= MLP_{sc}(V_{status\_att}, V_{global}) \end{aligned} \tag{12}$$

where $V_{status\_query} \in \mathbb{R}^{N\times D_{emb}}$, $V_{status\_att} \in \mathbb{R}^{1\times D_{emb}}$, and $V_{status} \in \mathbb{R}^{1\times D_{emb}}$.

We use a fully connected layer and sigmoid function to predict the above three types of labels $p_{object}$, $p_{action}$ and $p_{status}$ based on $V_{object}$, $V_{action}$, and $V_{status}$. Finally, we utilize the BCE loss function to supervise label prediction.

$$loss_c = -\sum_i BCE\_Loss(p_i, C_i) \tag{13}$$

where $i \in [object, action, status]$, and $C_i$ is the groundtruth of attribute $i$.

To summarize, based on the above multi-hierarchical crowd aware module, we finally capture the high-level and general global visual features $V_{global}$, local visual features $V_{local}$, and three types of visual features about crowd $V_{object}$, $V_{action}$, $V_{status}$.

### D. High-level Feature Fusion Module

Since the decoder generates the final sentence by predicting words step by step, it is unreasonable to send a fixed visual feature into the decoder to predict different words. Therefore, we perform feature fusion according to the hidden state of the decoder at the previous moment $h_{t-1}$ to obtain high-level visual features $V_h$, which can be dynamically adjusted according to $h_{t-1}$. $V_h$ is more suitable for the current state.

$$\begin{aligned} V &= [V_{global}, V_{object}, V_{action}, V_{status}]_{1st} \\ V' &= [W_1 h_{t-1}]_4 + W_2 V \\ \alpha &= softmax(W_3 V') \\ V_h &= \alpha V + AvgPool(V) \end{aligned} \tag{14}$$

where $W_*$ are trainable parameters, $[*]_4$ indicates the expanding operation by repeating features 4 times, and $[*,*,*,*]_{1st}$ represents the feature cascading operation on the first dimension. $V \in \mathbb{R}^{4\times D_{emb}}$,$V' \in \mathbb{R}^{4\times D_{emb}}$, $\alpha \in \mathbb{R}^{1\times 4}$, and $V_h \in \mathbb{R}^{1\times D_{emb}}$.

### E. Crowd Guided Decoder

In the decoding process, we adopt a simple LSTM with high-level crowd attribute features to generate a description. First, we input the previous word $w_{t-1}$, high-level crowd attribute feature $V_h$, global feature $V_{global}$, and previous crowd guided consistent feature $V_{t-1}^w$ into the LSTM.

$$\begin{aligned} V_{visual} &= V_h + V_{global} + V_{t-1}^w \\ V_{in} &= MLP([w_{t-1}, V_{visual}]_c) \\ (h_t, c_t) &= LSTM(V_{in}, (h_{t-1}, c_{t-1})) \end{aligned} \tag{15}$$

where $V_{t-1}^w$ is the semantic feature of the previous time, $w_{t-1}$ is the word prediction result, $V_{t-1}^\omega \in \mathbb{R}^{1\times D_{emb}}$, $V_{visual} \in \mathbb{R}^{1\times D_{emb}}$, $\omega_{t-1} \in \mathbb{R}^{1\times D_{words}}$, $V_{in} \in \mathbb{R}^{1\times (D_{emb}+D_{words})}$, $h_t \in \mathbb{R}^{1\times D_{hid}}$, and $c_t \in \mathbb{R}^{1\times D_{hid}}$.

In general, researchers use a simple MLP to directly decode the hidden status of LSTM $h_t$ to predict the $w_t$ of the sentence at the current moment, which ignores the importance of visual information in the process of word mapping. Moreover, loss of visual information accumulated over time also results in poor decoding results.

Since local visual features $V_{local}$ contain rich visual features of different regions, we use the hidden layer state $h_t$ of the LSTM to query the more representative visual features $V_{local}'$ in local visual features, and then utilize an MLP to extract a word prediction feature $V_t^w$ from $h_t$, $V_{local}'$ and $V_{in}$, which is consistent with the hidden layer state $h_t$ and local visual features $V_{local}$.

$$\begin{aligned} V' &= ReLU(W_1[h_t]_N \odot W_2 V_{local}) \\ \beta^1 &= softmax(W_3 V') \\ \beta^2 &= sigmoid(W_4 Avgpool(V')) \\ V_{local}' &= \beta^2 \odot (\beta^1 V_{local}) \\ V_t^w &= MLP([V_{local}', h_t, V_{in}]_c) \end{aligned} \tag{16}$$

where $W_*$ are trainable parameters, $V' \in \mathbb{R}^{N\times D_{hid}}$, $\beta^1 \in \mathbb{R}^{1\times N}$, $\beta^2 \in \mathbb{R}^{1\times D_{hid}}$, $V_{local}' \in \mathbb{R}^{1\times D_{hid}}$, and $V_t^\omega \in \mathbb{R}^{1\times D_{hid}}$, $D_{hid} = D_{emb}$.

After obtaining the crowd-guided visual features $V_t^w$, we finally use a fully connected layer to convert $V_t^w$ into the word vocabulary.

$$p_t = softmax(FC(tanh(V_t^w))) \tag{17}$$

where $p_t \in \mathbb{R}^{1\times N_{words}}$.

Similar to existing methods, we adopt cross-entropy loss to calculate the prediction loss of the captions:

$$loss_w = -\sum_{t=1}^{T} logP_t(p_t) \tag{18}$$

Finally, we define the final training loss as follows:

$$loss = loss_w + \beta \cdot loss_c \tag{19}$$

## V. Experiments

We conduct sufficient experiments with existing open-source methods on CrowdCaption and choose the classic caption method Up-Down [42] as our base method. In this section, we first introduce the training strategy, evaluation metrics and relevant parameter settings of the experiments, and then we analyze the impact of different modules. Finally, we show the experimental results from objective and subjective perspectives, respectively. The experimental results show that our network outperforms the state-of-the-art results and proves that the research on crowd attributes has benefits.

### A. Training Strategy

Following the standard and general methods of image captioning, we adopt a two-stage training strategy. In the first stage, we train our network by minimizing the cross-entropy loss and attribute class loss in Eq.19. In the second stage, we further use the idea of self-critical reward to optimize the network based on the CIDEr score.

### B. Evaluation Metrics

To evaluate the captions, we report commonly used evaluation metrics: BLEU@1–4 [44], METEOR [45], ROUGE-L [46], and CIDEr [47]. These are widely used to evaluate machine translation task and image captioning task, and can fully reflect the quality of captions. We use the Microsoft COCO server tools [48] as experimental evaluation tools. The values of evaluation metrics are expressed as percentages, which are positively correlated with the quality of generated captions.

### C. Experiment related settings

For the CrowdCaption dataset, we extract 36 bottom-up features and position information [42] for each image as visual information $F_r$ and $F_p$. $F_r$ is a 2048-dimensional vector, and $F_p$ contains the coordinates and score of each region with the structure of $[x; y; w; h; score]$. The number of regions in the image $N$ is 36. Taking into account the particularity of the crowd scenes, we exploit HRNet [43] to extract the feature spectrum of the human body pose as $F_c$, which contains rich information about the joint points of humans. The number of joints in human pose estimation $M$ is 34. The dimension of the human pose feature map for different parts is $128 \times 208$.

In the data preprocessing stage, we remove all stop words, words appearing fewer than two times, and finally collect 2660 words as the vocabulary for CrowdCaption. By analyzing and counting the caption annotations in CrowdCaption, we set 50 as the maximum length of the sentence. For all other state-of-the-art methods, we keep the parameters of the original network unchanged as much as possible, such as the dimension of the hidden layer and embedding layer, and we set the vocabulary and the length of the sentence corresponding to CrowdCaption as special experimental protocols.

For our method, the whole architecture is mainly implemented with Xmodaler [49] on Pytorch. For the baseline method, we choose the most classic method Up-Down [42] as our baseline method. In our method, the dimension of visual embedding is 1024, and the hidden layer dimension of LSTM is also 1024. What' more, the embedding dimension of words in the vocabulary is 512, and the balance parameter $\beta$ in the loss function is 0.2.

In the training stage, we first optimize with cross-entropy loss for 80 epochs, and then use the reward mechanism to optimize CIDEr score for 40 epochs. For the optimizer, we use the conventional Adam optimizer with a warm-up mechanism for 1000 iterations, and the initial learning rate of the network is 5e-4. When using the reward mechanism to optimize CIDEr score, we drop the learning rate by a factor of 10 and used linear decay for the learning rate every 3 epochs. In the testing stage, we choose the beam search method and set 3 as the beam size.

TABLE III: Ablation experiment of main components.

| MCA | HFF | CGD | B@4 | M | R | C |
|---|---|---|---|---|---|---|
| | | | 27.78 | 21.34 | 43.44 | 58.68 |
| ✓ | | | 28.53 | 21.75 | 44.29 | 61.81 |
| ✓ | ✓ | | 29.01 | 22.04 | 44.65 | 62.26 |
| ✓ | | ✓ | 29.23 | 21.92 | 44.62 | 63.63 |
| ✓ | ✓ | ✓ | 29.76 | 22.44 | 45.34 | 64.82 |

### D. Ablation Study

In order to fully analyze the effectiveness of the main modules and parameters, we conduct extensive ablation experiments on CrowdCaption.

**Effects of main components.** We first perform ablation studies on our proposed main components and show the results in Table-III. In Table-III, MCA represents the multi-hierarchical crowd aware module in the encoder, HFF represents the high-level feature fusion module, and CGD represents the crowd guided decoder. Compared with the baseline, the multi-hierarchical crowd aware (MCA) module produces an obvious improvement. In particular, the CIDEr increased by 3.1%, which indicates that the supplement of multi-hierarchical crowd attribute features greatly improves the richness and pertinence of visual information. When we use the high-level feature fusion (HFF) module to replace the average pooling, the results demonstrate the effectiveness of dynamic high-level visual features compared with fixed visual features. Dynamic high-level visual features can provide more crowd-targeted visual features for the decoder. The results of combining MCA and CGD show that relying only on the output of the hidden layer in the LSTM is insufficient, and

TABLE VII: Results of existing open source methods and our benchmark MAGC on CrowdCaption dataset. "CIDEr Score Optimization" means using the idea of self-critical reward in reinforcement learning to optimize CIDEr score during training process.

| Methods | Cross-Entropy Loss | | | | CIDEr Score Optimization | | | |
|---|---|---|---|---|---|---|---|---|
| | B@4 | M | R | C | B@4 | M | R | C |
| Show, attend and tell [24] | 26.98 | 20.78 | 42.97 | 56.51 | 28.29 | 21.47 | 44.71 | 61.78 |
| ConceptualCaptions [50] | 27.64 | 20.80 | 42.71 | 58.66 | 28.36 | 21.36 | 43.58 | 63.22 |
| Up-Down [42] | 27.78 | 21.34 | 43.44 | 58.68 | 29.70 | 22.01 | 45.04 | 64.79 |
| Meshed-Memory [51] | 28.32 | 21.28 | 43.18 | 59.22 | 29.33 | 21.55 | 43.74 | 63.89 |
| AoAnet [28] | 28.04 | 21.79 | 43.69 | 61.37 | 29.11 | 21.73 | 44.44 | 65.01 |
| X-LAN [33] | 28.91 | 21.75 | 43.89 | 62.07 | 29.59 | 22.06 | 44.69 | 66.07 |
| MAGC | 29.76 | 22.44 | 45.34 | 64.82 | 30.14 | 22.26 | 45.61 | 69.32 |

TABLE IV: Ablation experiment of visual features.

| Visual Features | B@4 | M | R | C |
|---|---|---|---|---|
| $F_r$+$F_p$+$F_c$ | 29.76 | 22.44 | 45.34 | 64.82 |
| $F_r$+$F_c$ | 28.98 | 21.95 | 44.64 | 63.69 |
| $F_r$+$F_p$ | 28.78 | 22.05 | 44.69 | 63.35 |
| $F_r$ | 28.64 | 21.87 | 44.61 | 62.71 |

TABLE V: Ablation experiment about number of layers in MCA.

| the number of layers | B@4 | M | R | C |
|---|---|---|---|---|
| 1 | 27.92 | 20.98 | 43.44 | 61.54 |
| 2 | 28.62 | 21.69 | 43.85 | 62.31 |
| 3 | 29.09 | 21.95 | 44.66 | 63.43 |
| 4 | 29.76 | 22.44 | 45.34 | 64.82 |

supplying more crowd visual information can promote a more accurate description.

**Effects of the visual features in multi-hierarchical structure.** Besides, we also conduct some ablation experiments to further study the improvement brought by different visual features in the multi-hierarchical crowd aware module. We separately consider the impact of different visual features, including region features $F_r$, position features $F_p$, and crowd posture features $F_c$. As shown in Table-IV, supplementing crowd posture information and position information can help the network generate better descriptions, especially in BLUE@4 and CIDEr. Performance improvements indicate that the position features and crowd posture information promote the enhancement of visual features, which can help the network capture more detailed information about the crowd and these features are useful in genetating precise descriptions.

**Effects of the hierarchy number of layers in MCA.** MCA represents the multi-hierarchical crowd aware module in the encoder. In order to further explore the effect of hierarchical structure on the caption generation process, we further conduct some ablation experiments on the number of layers of hierarchy in MCA. As shown in Table-V, as the number of layers increases, the network can generate higher quality descriptions for images. It can be clearly seen that the multi-hierarchical structure can help the network learn deeper visual information, thereby better helping the decoder fuse richer semantic features to generate more detailed captions.

TABLE VI: Ablation experiment about $\beta$ in loss function.

| hyperparameter $\beta$ | B@4 | M | R | C |
|---|---|---|---|---|
| 0.0 | 28.69 | 21.89 | 44.11 | 63.04 |
| 0.1 | 29.18 | 21.79 | 44.51 | 63.40 |
| 0.2 | 29.76 | 22.44 | 45.34 | 64.82 |
| 0.3 | 29.58 | 22.02 | 44.72 | 65.61 |
| 0.4 | 29.25 | 21.96 | 44.71 | 64.37 |
| 0.5 | 29.11 | 22.05 | 44.68 | 63.02 |

**Effects of attribute prediction loss.** Finally, we conduct ablation experiments on the importance of attribute prediction in the loss function. We set $\beta$ as 0.0, 0.1, 0.2, 0.3, 0.4, and 0.5. From a mathematical point of view, $\beta$ represents the importance between the precision of captions and the accuracy of attribute prediction. For a $\beta$ setting of 0.0, our method is only constrained by using word prediction loss during training. As shown in Table-VI, as the value of $\beta$ increases, the effectiveness of attribute prediction supervision is gradually reflected. The correct amount of supervisory information can assist the network in predicting more accurate words. However, when $\beta$ is greater than 0.4, the network pays more attention to attribute prediction than the quality of words. Considering all evaluation metrics, we set 0.2 as the hyperparameter for $\beta$.

### *E. Quantitative Results*

We use existing open-source methods to conduct a large number of experiments on the CrowdCaption dataset. Table-VII shows performance comparisons between the state-of-the-art (SoTA) models and our methods. In order to make fair comparisons, we train all methods based on their own strategies and parameter settings. According to the results in Table-VII, although existing approaches proposed in recent years are continuously improving, they still do not perform well on crowd scenes descriptions. As a classic caption method, Up-Down [42] achieves poor results in crowd

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3192729




Fig. 6: Visual examples of our method on CrowdCaption dataset, coupled with corresponding groundtruth annotations.

scenes, especially on the BLEU@4 and CIDEr metrics. X-LAN [33] is an excellent caption method in recent years, and its complex three-layer X-Linear attention block has still not achieved much improvement in crowd scenes. The main reason for the poor performance is that the caption method captures the general visual key point of descriptions only through the whole-image features from the encoder, which does not take special consideration for crowd scenes. When designating the crowd scene topic of images, the general feature extractions are insufficient and need to adjust for crowd scenes. Our benchmark method achieves 29.76%, 22.44%, 45.34% and 64.82% on BLEU@4, METEOR, ROUGE-L and CIDEr, respectively, which demonstrates the effectiveness of our method on crowd scenes. Based on the self-critical reward training strategy, the performance has significantly improved, especially ROUGE-L and CIDEr.

### *F. Qualitative Results*

In order to clearly evaluate the descriptions of crowd scenes, we visualize the qualitative results of our method in Figure-6. It can be observed that our method can generate detailed descriptions of a crowd's behavior and state. Especially in (b), although the crowd is not salient in the image, our method still describes the crowd on the beach and the state of the crowd under the umbrella, illustrating that our method is able to specifically understand the crowd and objects associated with the crowd in the image rather than the clearest salient object in the image. Some generated captions also prove that our method can generate detailed descriptions. For instance, in (c), our method can see the railing and crowd sitting behind the railing; in (d), very subtle photographic behavior can also be captured. In addition, our method is well described even for more difficult behavior, such as awarding awards. These subjective evaluation results also indicate the effectiveness of our method in crowd scenes.

## VI. Conclusion

In this paper, we proposed the crowd scenes caption dataset CrowdCaption for the first time. CrowdCaption focuses on assisting crowd scenes descriptions and provides comprehensive and diverse interrelated sentences for multiple perspectives, and all annotations contain a large amount of information about crowds and prominent crowd regions and objects, which differentiates it from other existing datasets. For crowd scenes image captioning, our benchmark method exploits multi-hierarchical crowd features in the encoder to obtain high-level crowd attribute features and further guide the decoding process of sentences to be more suitable crowd scenes. We finally achieved the state-of-the-art performance on CrowdCaption. We hope that the CrowdCaption dataset can promote the development of multimodal research related to crowd scenes understanding in the future.

## REFERENCES


[1] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[2] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.

[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[4] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: A dataset for image captioning with reading comprehension," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 742–758.

[5] A. P. Mathews, L. Xie, and X. He, "Senticap: Generating image descriptions with sentiments," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, pp. 3574–3580.

[6] N. Li, F. Chang, and C. Liu, "Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Transactions on Multimedia*, vol. 23, pp. 203–215, 2021.

[7] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 747–10 756.

[8] J. Zhang, L. Lin, J. Zhu, Y. Li, Y.-c. Chen, Y. Hu, and S. C. H. Hoi, "Attribute-aware pedestrian detection in a crowd," *IEEE Transactions on Multimedia*, vol. 23, pp. 3085–3097, 2021.

[9] X. Jiang, L. Zhang, T. Zhang, P. Lv, B. Zhou, Y. Pang, M. Xu, and C. Xu, "Density-aware multi-task learning for crowd counting," *IEEE Transactions on Multimedia*, vol. 23, pp. 443–453, 2021.

[10] X. Liu, G. Li, Z. Han, W. Zhang, Y. Yang, Q. Huang, and N. Sebe, "Exploiting sample correlation for crowd counting with multi-expert network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 3215–3224.

[11] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 863–10 872.

[12] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5385–5394.

[13] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets," *IEEE Transactions on Image Processing*, vol. 30, pp. 1439–1452, 2021.

[14] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.

[15] Q. Jia, H. Huang, and K. Q. Zhu, "Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 13 125–13 133.

[16] S. Liu, A. Zhang, Y. Li, J. Zhou, L. Xu, Z. Dong, and R. Zhang, "Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 2163–2171.

[17] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *CoRR*, vol. abs/1805.00123, 2018.

[18] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, and S. Lyu, "Detection, tracking, and counting meets drones in crowds: A benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7812–7821.

[19] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 74–93, 2017.

[20] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3337–3345.

[21] X. Yang, H. Zhang, D. Jin, Y. Liu, C. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in *Proceedings of the European Conference on Computer Vision*, vol. 12358, 2020, pp. 1–17.

[22] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8395–8404.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2048–2057.

[25] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3242–3250.

[26] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[27] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European Conference on Computer Vision*, vol. 11218. Springer, 2018, pp. 711–727.

[28] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4633–4642.

[29] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask learning for cross-domain image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2019.

[30] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder–decoder network for image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2942–2956, 2019.

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

[32] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9959–9968.

[33] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 968–10 977.

[34] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Exploring pairwise relationships adaptively from linguistic context in image captioning," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[35] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[36] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Integrating part of speech guidance for image captioning," *IEEE Transactions on Multimedia*, vol. 23, pp. 92–104, 2021.

[37] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. A. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 695–10 704.

[38] L. Yang, H. Wang, P. Tang, and Q. Li, "Captionnet: A tailor-made recurrent neural network for generating image descriptions," *IEEE Transactions on Multimedia*, vol. 23, pp. 835–845, 2021.

[39] L. Guo, J. Liu, S. Lu, and H. Lu, "Show, tell, and polish: Ruminant decoding for image captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2149–2162, 2020.

[40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[41] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 91–99.

[42] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the*

*IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[45] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014, pp. 376–380.

[46] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[47] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[48] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *Computer Science*, vol. abs/1504.00325, 2015.

[49] Y. Li, Y. Pan, J. Chen, T. Yao, and T. Mei, "X-modaler: A versatile and high-performance codebase for cross-modal analytics," in *Proceedings of the 29th ACM international conference on Multimedia*, 2021.

[50] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2556–2565.

[51] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 575–10 584.