Received October 5, 2021, accepted November 11, 2021, date of publication November 22, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129782

A Scientometric Visualization Analysis of Image Captioning Research From 2010 to 2020

WENXUAN LIU[1], HUAYI WU[1], KAI HU[2], QING LUO[3], AND XIAOQIANG CHENG[4]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan 430079, China
[2]Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Wuxi 214122, China
[3]School of Mathematics and Physics, Wuhan Institute of Technology, Wuhan 430205, China
[4]Faculty of Resources and Environmental Science, Hubei University, Wuhan 430062, China

Corresponding author: Huayi Wu (wuhuayi@whu.edu.cn)

This work was supported by the National Natural Science Foundation of China Program under Grant 41930107.

ABSTRACT Image captioning has gradually gained attention in the field of artificial intelligence and become an interesting and challenging task for image understanding. It needs to identify important objects in images, extract attributes, tell relationships, and help the machine generate human-like descriptions. Recent works in deep neural networks have greatly improved the performance of image caption models. However, machines are still unable to imitate the way humans think, talk and communicate, so image captioning remains an ongoing task. It is thus very important to keep up with the latest research and results in the field of image captioning whereas publications on this topic are numerous. Our work aims to help researchers to have a macro-level understanding of image captioning from four aspects: spatial-temporal distribution characteristics, collaborative networks, trends in subject research, and historical evolutionary path. We employ scientometric visualization methods to achieve this goal. The results show that China has published the largest amount of publications in image captioning, but the United States has the greatest impact on research in this area. Besides, thirteen academic groups are identified in the field of image description, with institutions such as Microsoft, Google, Australian National University, and Georgia Institute of Technology being the most prominent research institutions. Meanwhile, we find that evaluation methods, datasets, novel image captioning models based on generative adversarial networks, reinforcement learning, and Transformer, as well as remote sensing image captioning, are the new research trends. Lastly, we conclude that image captioning research has gone through three major development stages from 2010 to 2020, and on this basis, we propose a more comprehensive taxonomy of image captioning.

INDEX TERMS Image captioning, image description generation, scientometric analysis, visualization.

I. INTRODUCTION

As the representative technology of artificial intelligence (AI), deep learning has developed rapidly in recent years, and has been widely used throughout the fields of computer vision (CV) and natural language processing (NLP). Image captioning (or image caption) is an important part of image understanding, which could automatically generate human-like sentences for the given image [1]. This task requires the machine to be able to recognize objects in the image, understand the relationships between them, and express the main information by some concise natural language descriptions. Image captioning has been implemented extensively in the field of social media [2], remote sensing [3], robotics [4] and medical image report generation [5], it helps machine ''see'' the content of pictures, promotes machine intelligence, and will assist machines ''think'', ''talk'' and ''behave'' like humans in the future.

In language-vision community, image captioning has emerged as a popular research area that combines image understanding process and turning the image information into a natural language description [6]–[8]. From a CV point of view, image caption is a more challenging task than image recognition and image classification, owing to the extra challenge of recognizing the objects and actions within the image and making a meaningful description supported the contents found. When the computer encounters an image and outputs the corresponding visual context, it may describe features of the image (e.g., shape, color and texture), it can present the primary objects of the scenario, and even predict a

The associate editor coordinating the review of this manuscript and approving it for publication was Weipeng Jing.

dynamic relationship between people and objects (e.g., a man is playing frisbee game with a puppy). Furthermore, an image description can show objects that are not emerged, and tell a story beyond the visual content(e.g., dad drive his daughter to buy a gift for her mother, even though the picture only shows the dad is driving a car with his daughter, but does not contain the mother and the gift store). In short, image captioning technology requires not only accurate recognition of the image objects, but also contextual and background knowledge to understand the intent of the image expression.

Image description research has made a number of breakthroughs in the last decade. Vinyals *et al.* [9] extracted features from images and fed them into an recurrent neural network (RNN) with manually annotated statements to obtain image content descriptions. Xu *et al.* [10] combined long-short term memory (LSTM) models with attentional mechanisms in human vision to focus on salient objects when generating corresponding words. Shi and Zou [11] proposed a framework for remote sensing image description using convolutional neural network (CNN). Wu *et al.* [12] proposed incorporating high-level concepts into the CNN-RNN approach, and it achieved a significant improvement in image captioning. Lu *et al.* [13] introduced a visual ''sentinel'' strategy and designed an adaptive visual attention model. Qu *et al.* [14] proposed a deep multimodal neural network model for the semantic understanding of high-resolution remote sensing images. Lu *et al.* [15] constructed a large-scale aerial image dataset for the remote sensing image captioning problem. Yang *et al.* [16] proposed a multitask learning algorithm for cross-domain image captioning, which simultaneously optimized two coupled objectives of image captioning and text-to-image synthesis through a dual learning mechanism to improve the performance of image captioning. Deep neural networks help a lot to handle challenges in image caption field. To date, researchers have presented a number of review papers [17]–[21] to summarize the development of image description techniques.

Although these survey articles have provided a good literature review of image captioning, only a portion of the papers on visual captioning can be covered because re-searchers are generally not able to survey the complete data of the publications. Moreover, these review papers tend to focus on models, datasets, and evaluation methods, neglecting to explore the spatial-temporal distribution characteristics, research hotspots, and research communities in the development of the image captioning field. In order to grasp the development direction of image captioning technology from a macro perspective, and to help researchers gain a comprehensive understanding of the development status of the field, we propose a review method for image description based on scientometric analysis.

In this paper, we use scientometric analysis (or bibliometric analysis) methods to establish a systematic review of image captioning research. Being different from traditional interpretive reviews, the scientometric methods [22], [23] are data-driven approaches based on bibliographic data, and these approaches provide overall knowledge of research fields that scholars are interested in. The methods often employ particular metrics, e.g., co-word analysis [24], co-authorship analysis [25], and co-citation analysis [26] to visualize literatures.

The remaining of this paper is organized in four parts. Section II tells how the bibliographic data were collected and which methods we would use to do the analysis. Section III presents and expresses results found out by conducting bibliographic analysis. Section IV discusses results obtained through data-driven bibliographic analysis employed in this paper, and presents a taxonomy of image captioning approaches. Section V gives a summary that briefly answers:

(**Q1**) In image captioning field, what are the main research communities, what role do they play in the development of image captioning?

(**Q2**) Based on scientometric analysis, what are the trends and challenges in image captioning domain?

(**Q3**) What is the developing path of image captioning technology?

II. DATA AND METHODOLOGY

In this research, we provide an overall scientometric analysis framework for image captioning, including two stages of data preparation and bibliometric data analysis. As shown in Fig. 1, firstly, we collected data by setting specific conditions, and then we employed basic metrics, core research community mining, key topics and references identification and evolutionary path of image captioning to answer Q1-Q3. Main bibliographic methods, such as co-authorship analysis, co-occurrences analysis, and co-citation analysis, are mainly employed for the scientometric study of image captioning.

A. DATA PREPARING

There are several bibliographic data sources which can be applied in scientometric analysis. These include the abstract and citation index databases such as Web of science (WOS) and Scopus, the full text data bases such as ScienceDirect, SpringerLink, and ProQuest, the free online database sources such as Google Scholar, Microsoft Academic, Dimensions, and PubMed, and other data sources including PatentDerwent innovations index, BOOK Citation index. This paper uses the WOS to collect bibliographic data to do scientometric analysis of image captioning research, because the WOS contains widely accepted indices such as Science Citation Index Expanded (SCIE), and Social Science Citation Index (SSCI). We set the search conditions by restricting the database to the SCIE and SSCI indices. Considering many high quality papers are collected in conferences of Computer Science, we also picked the Conference Proceedings Citation Index-Science (CPCI-S) index. We confined the time span between 2010 and 2020, and set the article type to ''Article'' and the language type as ''English''. We conducted a topic word search using the terms ''image captioning'' or ''image caption''. Full conditions for collecting bibliographic data are presented in Table 1.

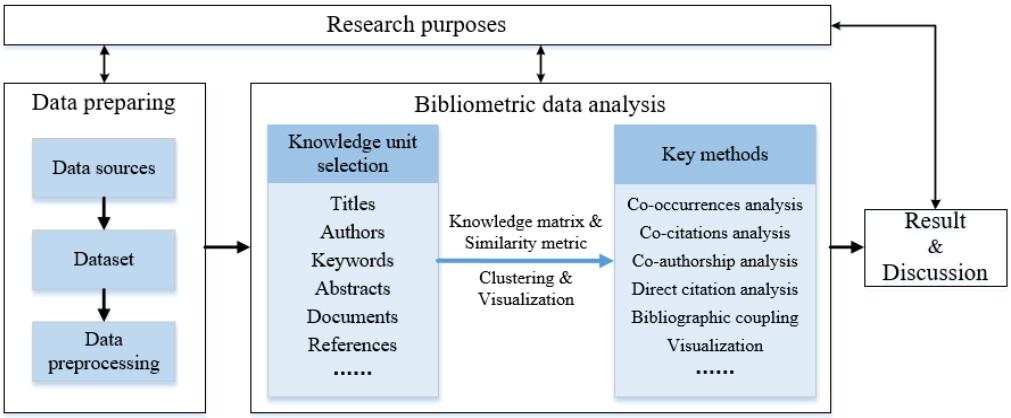


FIGURE 1. The scientometric analysis framework for image captioning.

TABLE 1. Bibliography searching conditions in WOS.

Topic	Timespan	Indexes	Document types	Language Type
"image captioning" or "image caption"	2010-2020	SCIE, SSCI, CPCI-S	Article	English

In a word, 697 papers were returned by these settings, including 235 journal articles and 462 proceedings articles. Unlike other disciplines, image captioning techniques are mostly published in conferences related to computer vision and natural language processing. The majorities of these papers were published after 2014 which indicates that image captioning research is emerging.

B. METHODS AND TOOLS

Several scientometric methods and softwares were employed to illustrate or visualize the research situation and progress of image captioning. Before answering Q1-Q3, it is helpful to have an overall impression of image captioning research.

1) BASIC STATISTICS

Basic statistics of scientometric analysis, e.g., yearly publication output, core journals/conferences, countries and institutions reported in this paper can provide "common-sense-knowledge" of a research domain. Besides, we also used Total Local Citation Score (TLCS) and Total Global Citation Score (TGCS) to indicate the influence of a scholar publication or organization. The TLCS is the citation counts of a journal/institution/country within the 697 papers, it reflects specialty in a specific research domain. Whereas the TGCS is the citation counts of a journal/institution/country within papers in WOS, it reflects global impacts. Software used in this analysis was Histcite [27] and VOSviewer [28].

2) ACADEMIC COMMUNITY MINING

It may not be enough for a scholar to have the overall picture of his/her research domain, because academic activities are with collaborations between individual scholars as well as institutions. Knowing academic communities are necessary for a researcher to develop his/her career. Fortunately, co-author analysis is a powerful method to explore academic communities. Co-author analysis means that two scholars co-authored one or more papers. VOSviewer is a software that can execute this method and display co-author networks, and thus find out academic communities.

3) VISUALIZATION METHODS OF HOT TOPICS, RESEARCH TRENDS AND EVOLUTIONARY PATH

It is often difficult for a beginner to make a traditional literature review, because discussions of hot topic, especially the research trends are usually based on experiences of a scholar who has ploughed in the fields for years. Nevertheless, scientometric analysis may provide a tool to do this job. Co-word or co-occurrence analysis is such a tool. Co-word means two words appearing in the same papers, abstracts or keywords units [29]. Co-word analysis often used to discover hot topics and research trends [30]. In addition, the document co-citation network can tell us the research front and knowledge structure in a scientific way. VOSviewer, CiteSpace [31] and HistCite are employed to do the visualization.

III. RESULTS AND ANALYSIS

A. BASIC BIBLIOMETRIC ANALYSIS

Bibliometric is the quantitative analysis of scholarly publications with the aim of demonstrating their impact on academic fields, it could estimate how much influence a selected research article has on future research. In this section, to obtain yearly output, the sources, spatial distribution, and main research institutions based on publications of image caption, the basic statistical analysis for image captioning research is established.

TABLE 2. The sources (conferences and journals) about image captioning based on number of publications.

(a) Top 10 conferences on image captioning research.

ID	Conferences	Recs	TLCS
1	IEEE Conference on Computer Vision and Pattern Recognition(CVPR)	68	1126
2	IEEE International Conference on Computer Vision(ICCV)	36	270
3	AAAI Conference on Artificial Intelligence(AAAI)	28	82
4	ACM Multimedia(MM)	20	37
5	European Conference on Computer Vision(ECCV)	18	91
6	International Conference on Image Processing(ICIP)	18	12
7	Neural Information Processing Systems(NeurIPS)	18	7
8	Annual Meeting of the Association for Computational Linguistics(ACL)	17	5
9	International Conference on Multimedia and Expo(ICME)	16	6
10	International Conference on Acoustics, Speech and Signal Processing(ICASSP)	8	1

(b) Top 10 journals on image captioning research.

ID	Journals	Recs	TLCS
1	IEEE Access	18	8
2	Neurocomputing	18	38
3	IEEE Transactions on Multimedia	16	42
4	Multimedia Tools and Applications	16	7
5	Applied Sciences-basel	12	0
6	ACM Transactions on Multimedia Computing Communications and Applications	11	0
7	IEEE Transactions on Image Processing	10	24
8	IEEE Transactions on Pattern Analysis And Machine Intelligence	8	161
9	Neural Processing Letters	8	5
10	Computer Vision and Image Understanding	6	5

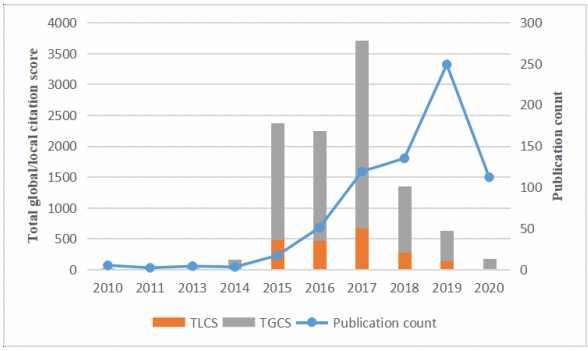


FIGURE 2. Yearly publication of image captioning research.

1) YEARLY PUBLICATION OUTPUT

The time series of the number of academic papers can be used as an important indicator to measure the research in a certain field, and the time distribution of the number of publications can directly reflect the research intensity of the field in different periods. In this paper, we use HistCite software to statistically analyze the number of papers detected by WOS for each year from 2010-2020, and plot the change curve of the number of papers published in image captioning research in different years. In Fig. 2, we can tell that the number of publications was no more than ten per year from 2010 to 2014. Then, the number of articles increased to 51 in 2016, and the number of papers published in each of the last four years has exceeded 100, of which 249 papers were published in 2019. Therefore, we can conclude that "Image Captioning" research is an emerging trend. The graph also shows that the TLCS and TGCS were largest in 2017, because there is a lag in literature citations, and researchers published a total of 119 articles in 2017, more than the previous decade combined.

2) THE SOURCES OF IMAGE CAPTION RESEARCH

From the analysis of HistCite software, the important conference and journal papers in the field of image captioning are shown in TABLE 2. More research papers related to image captioning are published in computer vision and artificial intelligence conferences such as CVPR, ICCV and AAAI than in journal papers. As can be seen from TABLE 2 (a), most of the research results in the field of image captioning are published in computer vision conferences represented by CVPR, ICCV, and ECCV, with CVPR being the most popular in terms of number and citations. Since ICCV and ECCV are held every two years, and CVPR is held once a year, the volume of publications is obviously more in CVPR. Meanwhile, image captioning research spans several fields such as image processing, natural language processing and speech recognition, so many researchers choose conferences such as ACM MM, ACL and ICASSP to present their research results. From TABLE 2 (b), it can be seen that the journals with the highest volume of image captioning studies are IEEE Access and Neurocomputing.

TABLE 3. Top 10 countries of image captioning from 2010 to 2020.

ID	Country	Recs	TLCS	TGCS
1	China	**338**	606	2317
2	USA	189	**1375**	**6215**
3	Australia	38	329	814
4	UK	33	70	318
5	India	31	4	24
6	South Korea	30	20	73
7	Germany	25	30	153
8	Japan	25	3	76
9	Singapore	21	103	515
10	Canada	16	59	272

Quantitative bibliometric analysis provides an objective description of the development of a field. Researchers in many disciplines may argue that journal papers have a greater influence on disciplinary development, but the trajectory of the last decade in the field of image captioning shows that many important research findings have been published in computer vision and artificial intelligence top conferences such as CVPR, ICCV and AAAI. As can be seen from Table 2, conference papers exceed journal papers in terms of both volume and impact on image captioning field from 2010 to 2020.

From a temporal perspective, the majority of journal papers on image captioning research were published after 2017. Compared to journal papers, researchers published 15 conference papers in 2015, and the number of conference papers published in each of the next five years increased much more than the number of journal papers. This means, with the development of image captioning, more and more researchers are joining this field of research. Since many of the papers published in international conferences such as CVPR, ICCV and ECCV are open source, researchers can more easily conduct innovative research or expand their research. Furthermore, due to the timeliness and innovation requirements of computer vision conferences, some researchers choose to publish their work in relevant journals.

Although the fruits of image captioning research are mostly submitted to international conferences, many scholars also prefer to publish their findings in journals. The possible reasons for this are mainly the followings. Firstly, journals usually have longer page limits. If a paper has too many experimental results to fit in a conference publication, then a journal affords an opportunity for inclusion. And review papers are usually published in journals due to their length. Secondly, despite the longer review cycle, journal reviews may be more detailed. Thirdly, researchers often prefer journals due to issues such as time, personal preference, university requirements, or practical needs.

3) SPATIAL DISTRIBUTION OF PUBLICATIONS BASED ON COUNTRIES

Table 3 shows the country distribution of the total number of publications and citations. HistCite software was run according to the pre-defined parameters to obtain the statistics of the number of publications and citations for the top 10 countries, and to generate a table of the distribution of publications and citations for image captioning studies by country. We can see that China and USA produced many of the articles and citations with image captioning, more than 75% of the total publications. Although China has the highest number of publications in this field with 338 articles, papers published in the United States are more influential, with 6215 global citations. This may be because much of the groundwork in artificial intelligence is done in the United States. In addition, from the number of publications in Table 3, we can see that Asian countries are more active in this field.

4) MAIN RESEARCH INSTITUTIONS

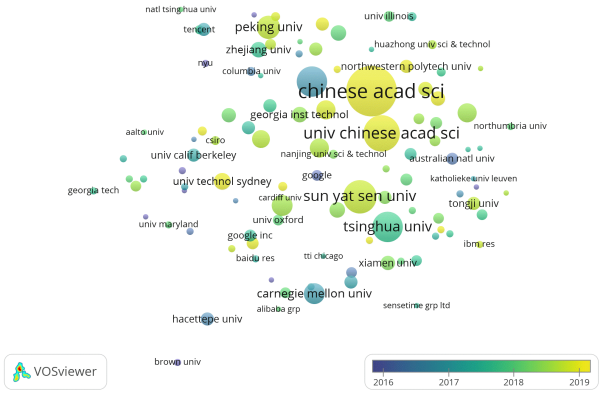
In order to obtain the core institutions in image captions generation field, data analysis is performed through HistCite and Vosviewer software on the knowledge graphs of research institutions, with dynamics that are often considered cutting-edge leaders in the field. We collected 20 institutes that focused on image captioning and published their re-search fruits at conferences or journals. From Table 4, we can tell that Chinese Academy of Sciences has the highest number of publications in the image captioning field, with 52 papers. However, the USA research institutions such as Microsoft Research, Facebook AI Research, and Google have gained significant citations and influence for their fundamental innovative work on image captioning. For example, Stanford University had only three articles, but it has 378 global citations. From the perspective of the volume of publications, although China has taken the lead in international competition in image captioning domain, the analysis of the impact of the research literature shows that China needs to strengthen its innovative research in the basic areas of artificial intelligence.

For HistCite software, an article with a high TGCS indicates that it has received more attention from scientists around the world. However, if an article has a high TGCS and a small TLCS, it means that this attention is mainly from scientists in other fields rather than image captioning. For beginners in the domain of image description generation, TLCS is more important. TLCS can help researchers to quickly locate classical literature in the field of image caption. As shown in Table 4, papers published by research institutions such as Google, Australian National University and Microsoft are classic literature in this field and worthy of in-depth learning by beginners.

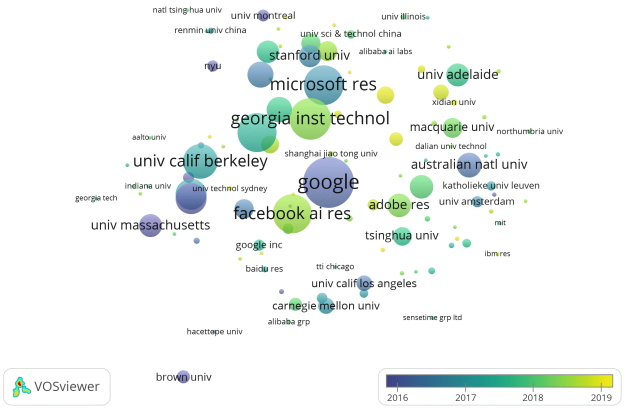
In addition, in terms of literature citations in the field of image description generation, prestigious universities in the traditional engineering domains (e.g., MIT, CMU, GIT and Stanford University) and Internet giants (e.g., Microsoft, Google and Facebook) in the United States dominated the highly cited papers. By analyzing the 697 publications, the VOSviewer software is used to derive Figure 3. The size of the labels represents the number of papers or citations. The bottom right corner of the visualization shows the color

TABLE 4. Selected 20 institutions of image captioning research based on volume of publications and citations.

ID	Institutions	Recs	TLCS	TGCS
1	Chinese Acad Sci	**52**	82	254
2	Univ Chinese Acad Sci	32	57	184
3	Sun Yat Sen Univ	28	59	132
4	Microsoft Res	24	185	887
5	Tsinghua Univ	24	67	282
6	Peking Univ	16	14	29
7	Carnegie Mellon Univ	14	48	225
8	Facebook AI Res	14	52	867
9	Beijing Univ Posts & Telecom	12	8	35
10	Nanyang Technol Univ	12	29	72
11	Shanghai Jiao Tong Univ	12	9	46
12	Univ Elect Sci & Technol China	11	26	262
13	Univ Technol Sydney	10	16	35
14	Beihang Univ	9	22	54
15	Georgia Inst Technol	9	106	931
16	Adobe Res	7	194	403
17	Natl Univ Singapore	7	82	452
18	Australian Natl Univ	6	213	436
19	Univ Rochester	6	176	402
20	Google	5	**396**	**1280**



(a) Document-based graph of image captioning research institutions.



(b) Citation-based graph of image captioning research institutions.

FIGURE 3. Graphs of image captioning research institutions.

bar from 2016 to 2019, illustrating the span of time the institution's papers have been cited. Echoing the findings in Table 4, Fig. 3(a) illustrates that Chinese research institutions, represented by the Chinese Academy of Sciences system, have invested a lot of effort in the field of image caption and have published several research papers. Fig. 3(b) demonstrates that the overlay visualization of citation-based map of image captioning research institutions. Take Google as an example, as the most highly cited paper in the field published in 2015, [9] has TLCS of 302. This article is considered to be an early pioneer in doing the image caption task, and it achieves better results with an ingenious modification of the Encoder-Decoder structure.

B. CORE RESEARCH COMMUNITIES MINING

The discovery of implicit research communities from scientists' collaborative networks is of great importance to understand the collaboration and communication patterns of researchers. In order to study the issue in depth, VOSviewer software is designed to find ''co-occurrence clustering''. This indicates that two things appearing at the same time implies that they are related, and there are various types of such relationships, such as co-authorship and word co-occurrence. In this section, we use VOSviewer to find different types of groups based on the clustering of relationship strength and direction measures. Fig. 4 demonstrates that a more pronounced pattern of collaboration among researchers in the image caption domain. We selected the main research groups, and the most linked collaborators in each group, to analyze the characteristics of image caption communities.

1) MICROSOFT-CENTRIC COMMUNITIES

In the 2015 MS COCO Image Captioning Challenge, Microsoft and Google tied for first place and the two separate systems performed equally well. In this competition, winners are determined based on two main metrics: The percentage of captions that are equal to or better than human-written captions, and the percentage of captions that pass the Turing Test. The campaign, based on a dataset provided by Microsoft [32], raised the popularity of image captioning and led to the inclusion of multiple research institutions in the study. This makes Microsoft one of the leaders in the field of image description. As shown in Fig. 5, since Microsoft Research has published a considerable number of papers which are highly cited on image captioning, the two communities (#1 and #4) resulting from our analysis are Microsoft-centric.

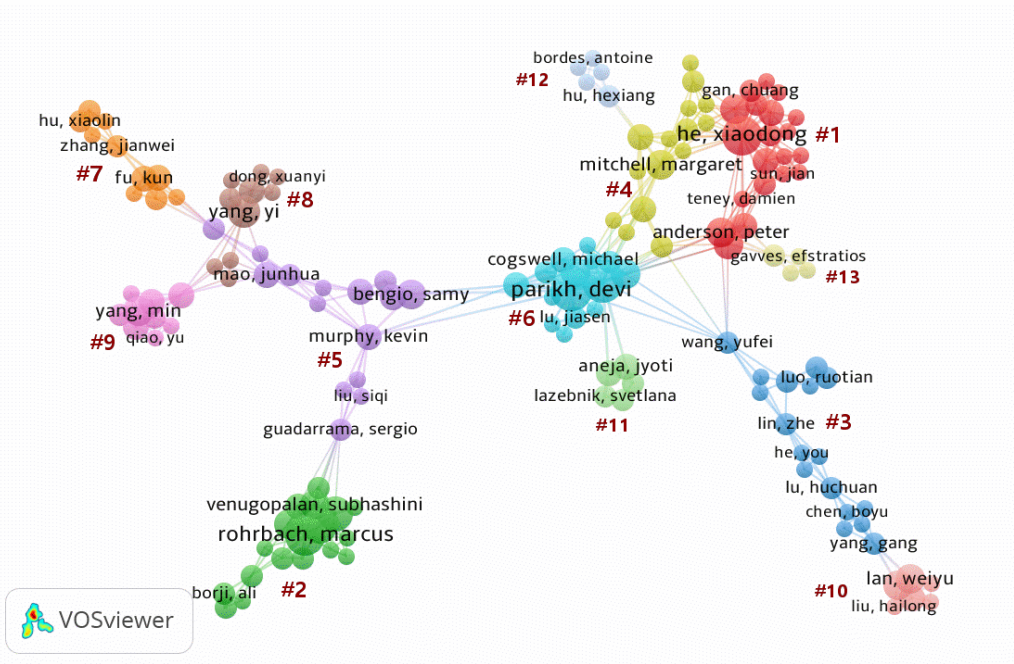
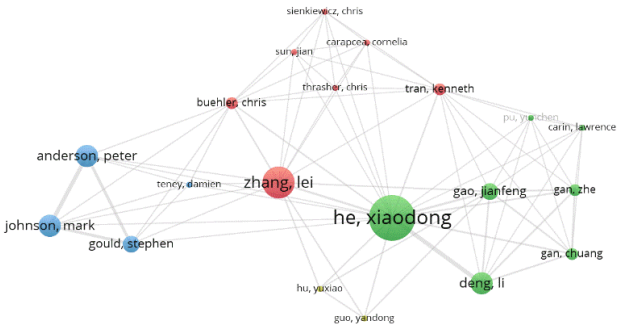
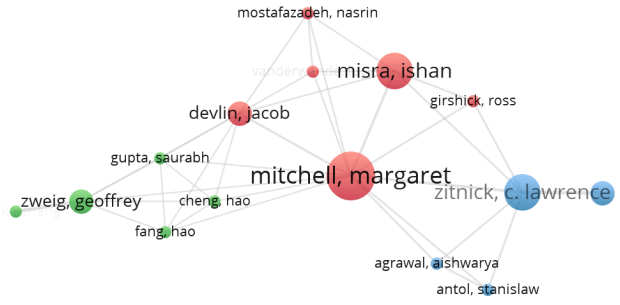


FIGURE 4. Co-author network of the publications in image captioning research from 2010 to 2020 (A graph-based map was obtained by analyzing 697 papers of image captions. This map, sorted by the number of authors in the community, has 13 clusters including 149 authors and 473 collaborative links.



(a) The collaborative network of cluster #1.



(b) The collaborative network of cluster #4.

FIGURE 5. Co-author network of Microsoft-centric communities.

Cluster #1 is the largest research community and includes the influential articles in the image captioning field. Many of the researchers in this community come from Microsoft Research. As shown in Fig. 5(a), the size of the nodes represents the degree of authorship. Larger nodes indicate that an author is more connected than the other authors represented by smaller nodes. Many of their articles have pioneered new research directions, such as semantic composition networks [33], bottom-up and top-down attention [34], and StyleNet [35]. Besides, Cluster #4 in Fig. 5(b) is also an academic community composed mainly of Microsoft scholars.

A. Peter *et al.* proposed a combined top-down and bottom-up attention model approach for application to problems related to visual scene understanding and visual question and answer (VQA) systems. The bottom-up attention model (generally using Faster R-CNN) is used to extract regions interest in images and obtain object features, while the top-down attention model is used to learn the weights corresponding to the features (generally using LSTM), so as to achieve a deeper understanding of visual images. This work has a total global citation of 191, and a number of studies have chosen this paper for experimental comparison. A. Peter proposed this method and refreshed the results of image captioning during his internship at Microsoft. His team won first place in the 2017 VQA Challenge. Moreover, A. Peter, J. Mark, and G. Stephen proposed a new evaluation method for image caption generation called SPICE [36].

Z. Gan *et al.* focused on semantic concept problem in image captioning, they proposed a method that used semantic information to integrate with recurrent neural network parameters. In order to generate attractive captions for images with different styles, C. Gan *et al.* propose a novel framework named StyleNet. This paper is the first to investigate the

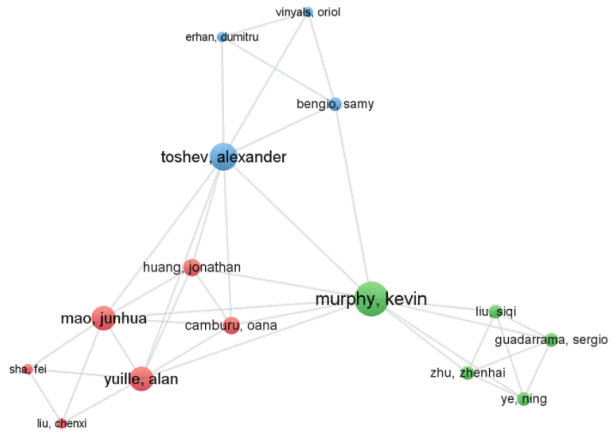


FIGURE 6. Co-author network of Google-centric research community.

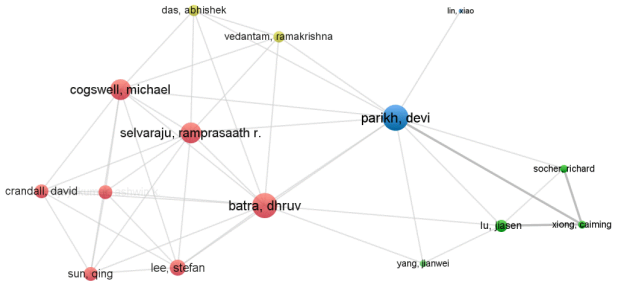


FIGURE 7. Co-author network of visualization-centric research community.

problem of using styles to generate attractive image captions without using supervised special image-caption pairing data.

Fang *et al.* [37] proposed a novel approach for automatically generating image descriptions. The methodology of this paper is divided into three main steps: visual detector, language model, and multimodal similarity model. It trained on images and corresponding captions, and learnt to extract nouns, verbs, and adjectives in the image. In another article from Microsoft Research, Devlin *et al.* [38] achieved state-of-the-art results in image captioning by using convolutional neural network and recurrent neural network.

2) GOOGLE-CENTRIC COMMUNITY

As shown in Fig. 6, Google is definitely one of the top research institutions. Despite the small number of its papers, it has not affected its influence in the field. As one of the pioneering papers in the domain of automatic visual description, Vinyals *et al.* [9] presented an end-to-end model to solve the image caption problem by introducing the Encoder-Decoder model to the field of Neural Image Captioning (NIC). The Encoder-Decoder architecture solved the problem of inconsistent length mapping and ushered in a new era of image captioning technology, leading a large number of scholars to conduct research in this area. In 2017, Vinyals *et al.* open sourced the latest version of the image description system on TensorFlow, and this public release contained significant improvements to the computer vision component of the image description system compared to the 2015 version, allowing for faster training and outputting more fine-grained, accurate descriptions [39].

J. H. Mao is one of the most active researchers in the community. In [40], Mao *et al.* proposed a multimodal Recurrent Neural Networks when he worked in Baidu Research, which has a deep convolutional network for images and a deep recurrent neural network for sentences. This approach is one of the important methods for image captioning studies using multimodal space. Moreover, when Mao was an intern at Google, he proposed a model that can generate an unambiguous caption of a specific object in description generation module, and can also comprehend an expression to infer object which is described [41].

3) VISUALIZATION-CENTRIC COMMUNITY

GIT (Georgia Institute of Technology) is one of the highly cited institutions, with nine publications in our dataset. From Fig. 7, the image caption visualization research community is composed of a number of highly cited authors from GIT. In perspective of the impact of research results, Selvaraju *et al.* [42] proposed a visual explanation method for CNN-based image captioning, which can visualize the features learned by the convolutional neural network, and help to understand how the neural network works and the decision-making process. This article is actually an evolved version of Class Activation Mapping (CAM), where the regions of interest in the model are distinguished by heat maps.

4) OTHER RESEARCH COMMUNITIES

Group #2 (in Fig. 4) has 18 researchers, the second largest community from the analysis result. As the biggest node in this community, M. Rohrbach published eight papers on image-to-text research, many of them related to video description. From Google scholar, he is a research scientist of Facebook AI research, and one of the highly cited authors in the image captioning field. T. Darrell is a highly cited author in the field of computer vision, has made outstanding contributions in the area of semantic segmentation of images, and is one of the co-authors of the deep learning framework Caffe. K. Saenko has an income-generating study of domain adaptation methods in machine learning. In [43], Donahue *et al.* designed a novel recurrent convolutional architecture of CNN-LSTM for behavior recognition, image caption, and video description. For automatic video to text research, S. Venugopalan, M. Rohrbach and J. Donahue *et al.* proposed the solution idea of using LSTM to solve the variable length of video and text, and the design of S2VT network structure for the whole video description are relatively classic [44].

As shown in Fig. 8, many researchers of cluster #3 use image captioning methods to solve other problems. H.C. Lu has the most links with other collaborators, his main research area is not image captioning, but visual tracking,

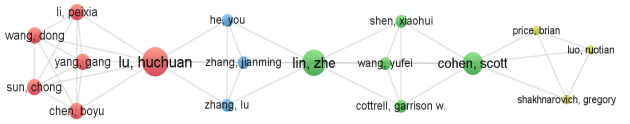


FIGURE 8. Co-author network of community #3.

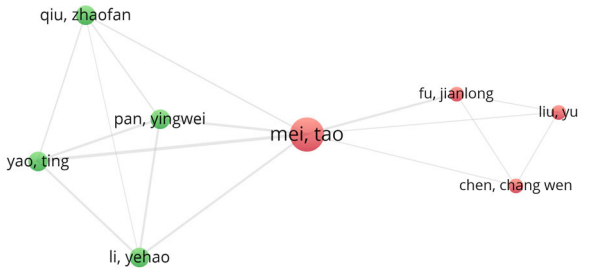


FIGURE 9. Co-author network of T. Mei-T. Yao *et al.*

object detection, etc. In [45], Zhang *et al.* designed a CapSal model to solve the problem of target detection in complex scenes, which consists of two sub-networks: image captioning network and local-global perception network. Most of the current research in image-to-text area focuses on captioning of single images, in [46], Tan *et al.* tried to generate relational captions for two images which is useful in various practical applications (e.g., image editing, difference interpretation, and retrieval). This paper opens up a new research direction in image captioning technology. Wang *et al.* [47] proposed that the process of human recognizing diagrams should be to locate the position of pictures and their relationship first, and then to elaborate the properties of objects. This article designed a coarse-to-fine approach based on this, first generating the skeleton sentence, then generating the corresponding attribute phrase, and finally synthesizing these two parts into a complete caption. S. Cohen is from Adobe Research, and has three papers with respect to image caption generation based on our dataset. In [48], Luo *et al.* focused on the problem of generating captions that are too general in visual language generation. In order to generate better captions, the idea of this article is to add discriminability as a training target during training process. In addition, Cohen *et al.* investigated the problem of figure caption generation where the goal is to automatically generate a natural language description for a given figure [49].

Fu *et al.* [50] from group #7 proposed an image captioning system that exploited the parallel structures between images and sentences. In [51], they proposed Image-Text Surgery approach to synthesize pseudoimage-sentence pairs, which can alleviate the expensive manpower of labeling data. In 2019, they presented a novel training objective for image captioning that consisted of two parts representing explicit and implicit knowledge respectively [52]. In cluster #8, Dong *et al.* [53] proposed a Fast Parameter Adaptation for Image-Text Modeling (FPAIT) that can jointly understand image and text data by a few examples. Wu *et al.* [54] introduced the zero-shot novel object captioning task, and proposed a Decoupled Novel Object Captioner (DNOC) framework that can fully decouple the language sequence model from the object descriptions. Researchers in group #9 focused on cross-domain image captioning [16], [55]. They used dual learning and multitask learning to improve the performance of image captioning system. W.Y. Lan and X.R. Li *et al.* from community #10 have mainly studied cross-lingual image captioning. In [56], they proposed a fluency-guided learning framework to learn a cross-lingual captioning model from machine-translated sentences. To enable image captioning applications that push the boundaries of languages, X.R. Li *et al.* proposed COCO-CN to enrich MS-COCO dataset with manually written Chinese sentences and tags [57]. Considering LSTM units were complex and inherently sequential across time, Aneja *et al.* [1] from group #11 developed a convolutional image captioning technique. In [58], Aneja *et al.* proposed SeqCVAE which learns a latent space for every word position. Shuster *et al.* [59] from cluster #12 proposed PERSONALITY-CAPTIONS, where the goal is to be as engaging to humans as possible by incorporating controllable style and personality traits. X. Jia *et al.* in group #13 presented an extension of the long short term memory (LSTM) model, which they coined gLSTM for short [60].

Since the scientometric approaches for mining collaborator networks relied on dataset as well as algorithms of visualization software, they could not cover all research communities to some extent. In Fig. 9, it shows an active research community of image captioning field. Yao *et al.* [7] investigated the effect of image attribute features on the description results, where image attribute features are extracted by a multi-instance learning approach. In [61], [62], T. Yao, T. Mei and Y.W. Pan *et al.* presented Long Short-Term Memory with Copying Mechanism (LSTM-C) and Long Short-Term Memory with Pointing (LSTM-P) to describe objects outside of training corpora (i.e., novel objects). The authors also studied the work of video captioning, and presented a new large-scale video description dataset called MSR-VTT [63], and they demonstrated a video captioning bot named Seeing Bot [64]. Besides, Yao *et al.* [65] presented a hierarchy parsing architecture which integrated hierarchical structure into image encoder to boost captioning.

C. KEY TOPIC AND REFERENCES IDENTIFICATION

Co-authoring networks can demonstrate the social links among scientists in the similar research community. However, some small research groups or independent authors cannot be revealed by co-author analysis, and some of them always have important research ideas. Term co-occurrence analysis could reveal important research topics in image captioning without considering co-authorship. For example, some people who are not co-authors of the same paper may use the same or similar key terms. Thus, the word co-occurrence analysis allows for greater capture of the information structure of textual topics.

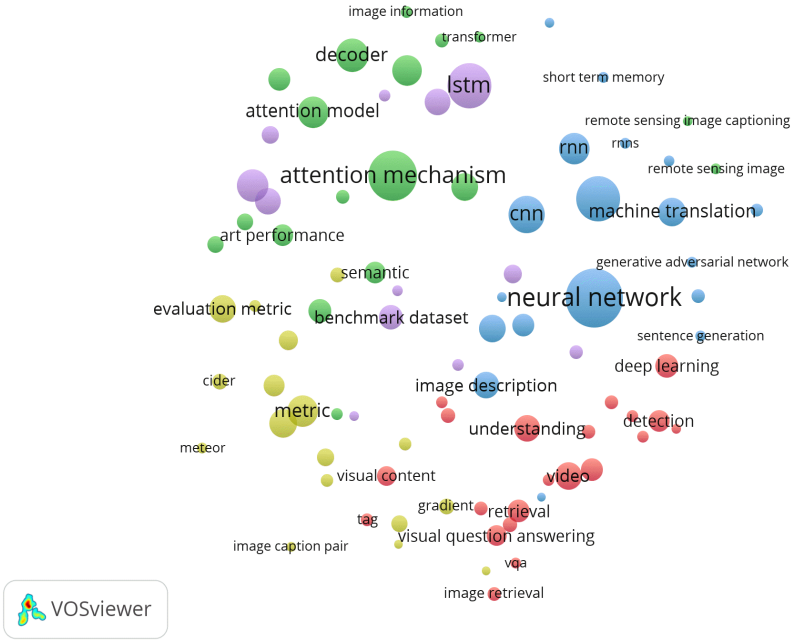


FIGURE 10. Term co-occurrence map of image captioning by VOSviewer (The size of each label indicates the number of occurrences and each color indicates the clustering of the different terms).

1) TERM CO-OCCURRENCE ANALYSIS

Keywords field (containing the title and keywords of the article) are the core summary of an essay and an analysis of keywords field in an essay can provide a glimpse into the topic of the essay. Whereas several keywords given in a paper must be related in some way, this association can be expressed in terms of the frequency of co-occurrence. It is generally accepted that the more frequent a word pair appears in the same document, the closer the relationship between the two topics is. In Fig. 10, analysis by VOSviewer shows that the main research architectures in the field of image captioning in the last decade, especially since the introduction of neural networks into the vision of computer vision. After Google proposed the neural image captioning in 2015, the Encoder-Decoder architecture was widely used and a large number of papers based on this architecture appeared. Later, after the attention mechanism was proposed, researchers combined the two and published a variety of image description generation related articles based on it. Therefore, the ''neural network'' and the ''attention mechanism'' are the largest labels in the graph.

The graph also reveals the dominant evaluation metrics and datasets in the field of image captioning. Currently, the criteria for evaluating the quality of automatic image caption can be divided into two categories: human evaluation and machine evaluation. Machine evaluation is fast and inexpensive, but far less accurate than human evaluation. From the leaderboard of the MS COCO competition, it can be seen that the commonly used metrics are BLEU [66], Meteor [67], [68], ROUGE [69], CIDEr [70] and SPICE [36]. The first two are judged for machine translation, the third for automatic abstraction, and the last two are supposed to be customized for image caption. The disadvantage of these evaluations is that they focus mainly on the shallow features of the sentences and only match the number of words, ignoring the deep semantics of the sentences. Hence, more accurate and efficient evaluation criteria need to be developed. Furthermore, as we can see, Flichr8k, Flichr30k, and MSCOCO are popular datasets in image captioning community.

As new directions and research areas, generative adversarial networks (GANs), reinforcement learning (RL), and Transformer methods are used for natural language generation in image caption field. Some of the current automatic image description methods adopt reinforcement learning and adversarial learning to annotate images, which can lead to more semantically rich text descriptions. These two methods help to implement unsupervised image captioning. The Transformer models has achieved state-of-the-art results in natural language processing tasks. For visual description tasks, Transformer models, which reduce structural complexity and explore scalability and training efficiency, are also emerging as a new research direction.

At the same time, remote sensing image captioning is also a new research direction that has been developed in recent years. Many of the published image description papers in remote sensing use CNN-RNN (or LSTM) models based on the attention mechanism. We can also note that visual

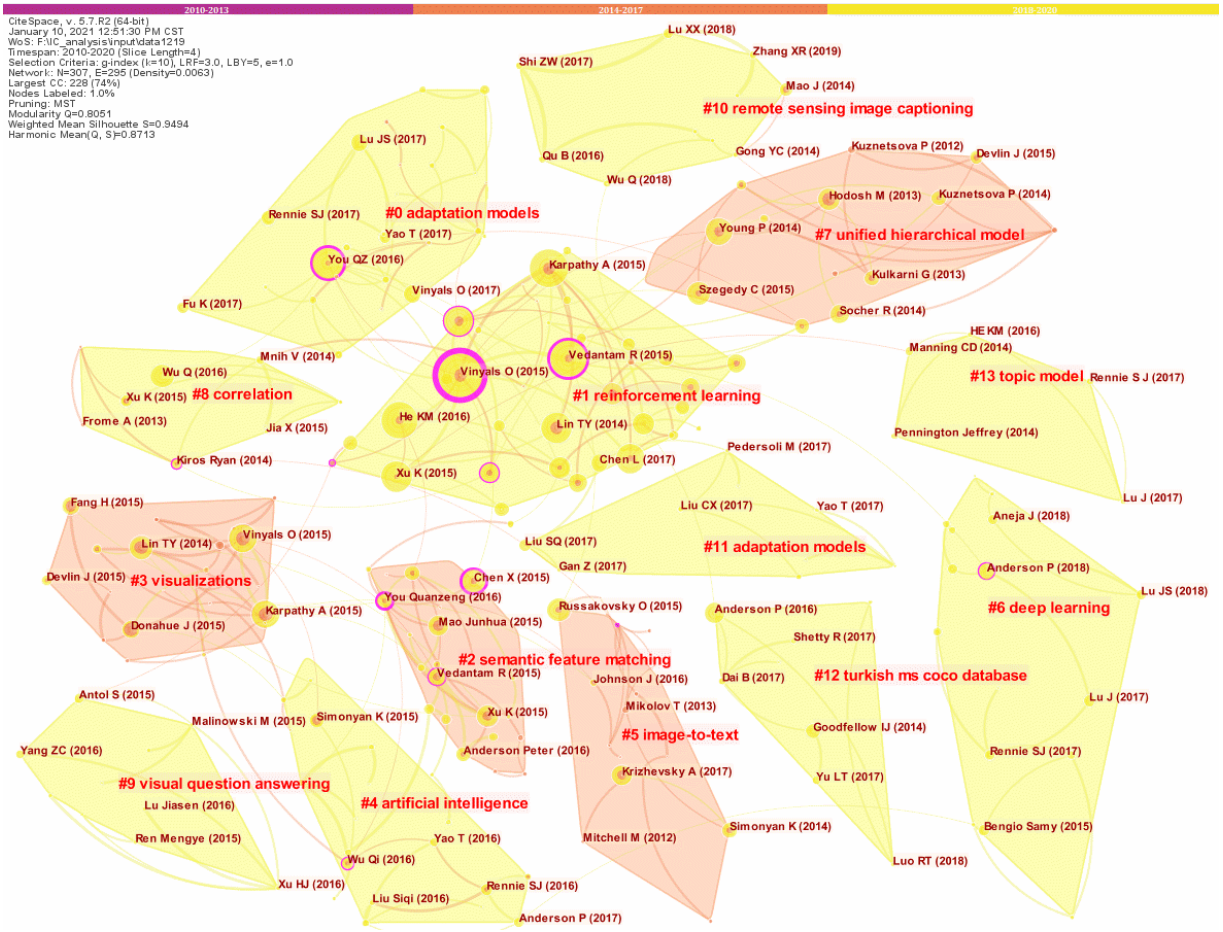


FIGURE 11. Co-citation based graph of image captioning research.

question answering (VQA), image retrieval, object detection and image captioning are among the more relevant subfields of computer vision and their development is in a mutually reinforcing relationship.

2) CO-CITATION ANALYSIS

Two (or more) papers that are simultaneously cited by one or more subsequent publications are said to constitute a co-citation relationship. Co-citation relationships in the literature change over time, and the study of co-citation networks in the literature allows the development and evolution of a discipline to be explored. For automatic captioning research, we use the CiteSpace clustering function to perform a cluster analysis of literature co-citations and mine similar literature for common themes, and the visualization results are shown in Fig. 11.

Firstly, we can get which articles are among the highly cited papers from the graph. The larger the node, the higher the frequency of co-citation is indicated. Also, the larger the purple outer circle, the higher the mediated centrality of the node. We can see that Vinyals *et al.* [9], Vedantam *et al.* [70], and You *et al.* [71] are very important papers in the field of image captioning. So cross-disciplinary experts and beginners can obtain the main results of image captioning from these co-cited high-frequency literatures. Secondly, we can also see from the graph which literatures are more closely associated. It means that these documents often appear together in multiple later publications, and the co-referenced documents are certainly similar in content. We use CiteSpace's cluster analysis method to uncover common themes in the related documents. From cluster #0 and cluster #11, we can conclude that domain adaptation model has emerged as a new learning technique to tackle the shortcomings of large amounts of labelled data. From cluster #2, #3, #5 and #7, in which the articles are published from early stage of image captioning technology, demonstrates that ''semantic feature matching'', ''visualization'', and ''unified hierarchical models'' are important in captioning field. Similarly, cluster #1 and #10 shows that ''reinforcement learning'' and ''remote sensing image captioning'' are the topics were extracted from the corresponding cited literature. Take group #10 as an example, Qu *et al.* [14], Shi and Zou [11], Lu *et al.* [15], and Zhang *et al.* [72] focused on remote sensing image captioning research. These papers are also cited

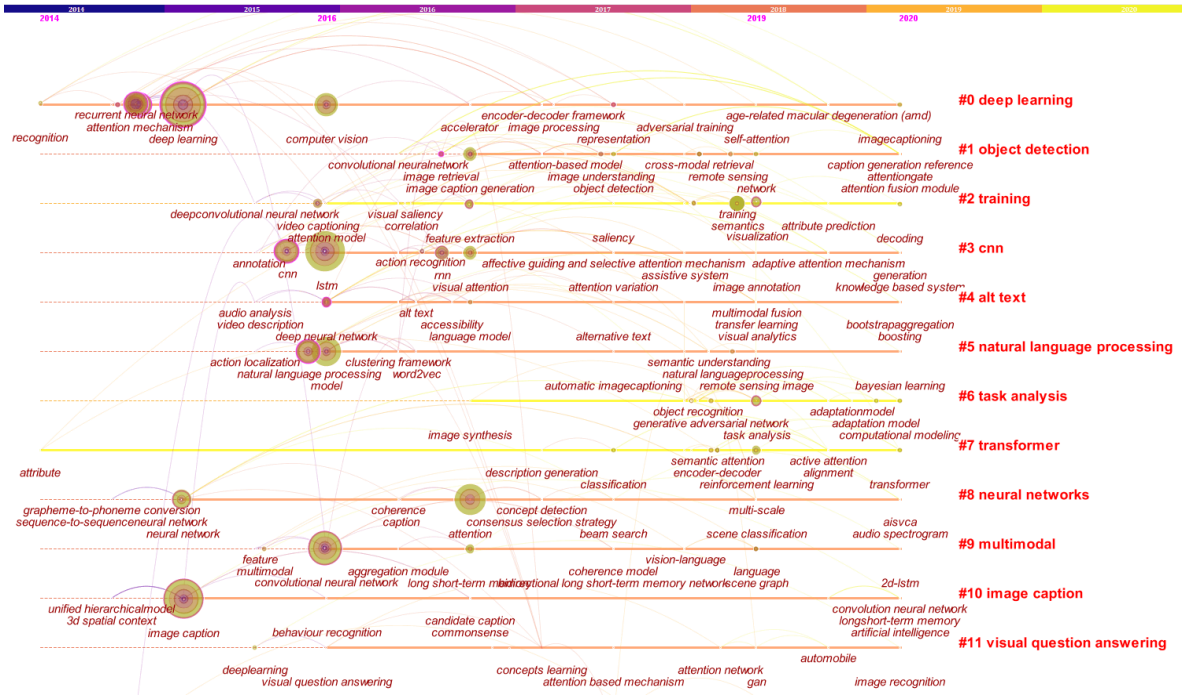


FIGURE 12. Timeline view of image captioning research from 2010 to 2020 by CiteSpace.

in the citation literature (although the other three articles are not related to remote sensing). So CiteSpace adopted ''remote sensing image captioning'' as the theme for this group. If researchers are interested in remote sensing image captioning techniques, then they could start by understanding the main innovation points of the above mentioned papers.

D. EVOLUTIONARY PATH OF IMAGE CAPTIONING

Historical reconstructions of scientific evolution can be depicted chronologically as the development of a network of citation relations extracted from the scientific literatures. We would like to illustrate the evolution of image captioning in recent years by combining the timeline view (Fig.12) generated by CiteSpace and the timeline based map (Fig.13) generated by HistCite.

Fig. 12 shows the development of keywords in 684 image captioning research papers, spanning the years 2014 to 2020. From top to bottom, the rightmost column presents the top 12 categories obtained by clustering all keywords (including IEEE keywords and author keywords) from largest to smallest, where ''task analysis'' and ''training'' are IEEE keywords. From left to right is the timeline of the development of image captioning techniques. The clustering results mainly included object detection, natural language processing, multimodal and visual question answering, as well as deep learning related methods such as ''neural networks'' and ''CNN''. It can be seen that Transformer has become one of the frontier research hotspot. From the visualization results of high frequency keywords, ''attention mechanism'', ''attention model'' and ''attention network'' and other attention-related techniques have become popular research directions after 2015. Meanwhile, the CNN-LSTM model has become one of the mainstream models used in research. In addition, image captioning techniques are gradually applied to ''remote sensing'', ''social media'', and ''medical image reporting''. For example, since automatic captioning tools have the potential to empower partially sighted people to know more about social media images without having to rely on human-authored alt text or asking a sighted person. Macleod *et al.* [90] provided the first evaluation of image captions generated by a full-sentence algorithm for the blind and visually impaired people.

Fig. 13 presents the citation network obtained after visualization by HistCite, which is drawn directly using the relationship between the cited literature and the references. It utilizes the highly cited literature in the field, and is temporal in nature. Form the research results of the most cited literature in each year, we can see that [9] presented the Neural Image Captioning model for the first time in 2015, bringing the Encoder-Decoder architecture to researchers and opening up a new era of image caption research. In 2016, [71] introduced semantic attention mechanisms into the study of image description, which combined visual information from top-down and bottom-up approaches in a complex neural network framework. This attention-based model is the most widely circulated of the many related methods. Then, [83] proposed a spatial and channel-wise attention based model in 2017. Reference [34] exploited bottom-up and top-down attention to improve the performance of image captioning and visual question and answer in 2018. In conclusion, the

TABLE 5. An overview of models, datasets, and evaluation metrics for the 35 most cited references.

ID	Publications	Vision Model	Language Model	Datasets	Evaluation Metrics	TLCS
16	X. Chen et al. 2015 [73]	VGGNet	LSTM	PASCAL 1K, Flicker8K/30K, MS COCO	BLEU, METEOR, CIDEr	48
17	O. Vinyals et al. 2015 [9]	GoogLeNet	LSTM	PASCAL VOC 2008, Flicker8k/30k, MS COCO, SUB	BLEU, METEOR, CIDEr, R@K, human evaluation	302
18	X. Jia et al. 2015 [60]	VGGNet	LSTM	Flicker8K/30K, MS COCO	BLEU, METEOR, CIDEr	69
19	J.H. Mao et al. 2015 [74]	AlexNet, VGGNet	RNN	NewObj-Cat, NewObj-Motor, NC-3	BLEU, METEOR	26
20	S. Vengopalan et al. 2015 [44]	CaffeNet, VGGNet	LSTM	MSVD, MPII-MD, M-VAD	METEOR	21
31	K. Cho et al. 2015 [75]	CNN	RNN	Flicker8k/30k, MS COCO	BLEU, CIDEr, METEOR, human evaluation	16
38	L.A. Hendricks et al. 2016 [76]	VGGNet	LSTM	MS COCO, ImageNet; CaptionTxt, WebCorpus	BLEU, METEOR, F1-score	17
39	J.H. Mao et al. 2016 [41]	VGGNet	LSTM	Google Refexp, UNC-Ref-COCO	BLEU, METEOR, CIDEr, human evaluation	14
40	Q. Wu et al. 2016 [12]	VGGNet	LSTM	Flicker8k/30k, MS COCO	BLEU, METEOR, CIDEr, sentence perplexity	79
44	J. Johnson et al. 2016 [77]	VGGNet	LSTM	Visual Genome	AP, IoU, METEOR	44
45	Q.Z. You et al. 2016 [71]	GoogLeNet	RNN	Flicker30k, MS COCO	BLEU, METEOR, ROUGE-L, CIDEr	174
60	P. Anderson et al. 2016 [36]	---	---	PASCAL-50S, Flicker8K, Composite Dataset, MS COCO	SPICE	77
62	H.J. Xu et al. 2016 [78]	SMem-VQA		DAQUAR, VQA	Accuracy	12
102	S.Q. Liu et al. 2017 [79]	Inception-v3	LSTM	MS COCO	SPIDEr, human evaluation	31
104	M. Pedersoli et al. 2017 [80]	VGGNet	RNN	MS COCO	BLEU, METEOR, CIDEr	14
107	B. Dai et al. 2017 [81]	VGGNet	LSTM	Flicker30k, MS COCO	BLEU, METEOR, ROUGE-L, CIDEr, SPICE, E-NGAN, E-GAN	25
111	T. Yao et al. 2017 [7]	GoogLeNet	LSTM	MS COCO	BLEU, METEOR, ROUGE-L, CIDEr, SPICE	31
128	Z. Gan et al. 2017 [33]	ResNet	LSTM	Flicker30k, MS COCO, Youtube2Text	BLEU, METEOR, ROUGE-L, CIDEr	12
131	S.J. Rennie et al. 2017 [82]	ResNet	LSTM	MS COCO	BLEU, METEOR, ROUGE-L, CIDEr	42
134	J.S. Lu et al. 2017 [13]	ResNet	LSTM	Flicker30k, MS COCO	BLEU, METEOR, CIDEr, ROUGE-L	67
139	T. Yao et al. 2017 [61]	VGGNet	LSTM	MS COCO, ImageNet	METEOR, F1-score, Accuracy, Novel	12
140	L. Chen et al. 2017 [83]	ResNet,VGGNet	LSTM	Flicker8k/30k, MS COCO	BLEU, METEOR, CIDEr, ROUGE-L	68
182	L.H. Li et al. 2017 [84]	VGGNet, Faster R-CNN	LSTM	MS COCO	BLEU, METEOR, CIDEr, ROUGE-L	16
183	C.X. Liu et al. 2017 [85]	VGGNet	LSTM	Flicker30k, MS COCO	BLEU, METEOR	13
188	O. Vinyals et al. 2017 [39]	GoogLeNet	LSTM	PASCAL VOC 2008, Flicker8k/30k, MS COCO, SUB	BLEU, METEOR, CIDEr, ROUGE, R@K human evaluation	60
189	A. Karpathy et al. 2017 [86]	VGGNet	RNN	Flicker8k/30k, MS COCO, Visual Genome	BLEU, METEOR, CIDEr, R@K	36
190	J. Donahue et al. 2017 [43]	CaffeNet, VGGNet	LSTM	Flicker30k, MS COCO	BLEU, METEOR, ROUGE-L, CIDEr	20
193	Z.W. Shi et al. 2017 [11]	VGGNet	LSTM	Google Earth images, GF-2 images	Precision-recall curves, confusion matrix	17
194	L.L. Gao et al. 2017 [87]	Inception-v3	LSTM	MSVD, MSR-VTT	BLEU, METEOR, CIDEr	13
200	K. Fu et al. 2017 [50]	ResNet	LSTM	Flicker8k/30k, MS COCO	BLEU, METEOR, ROUGE-L, CIDEr	28
235	J. Aneja et al. 2018 [1]	VGGNet	CNN model	MS COCO	BLEU, METEOR, ROUGE-L, CIDEr	23
238	P. Anderson et al. 2018 [34]	ResNet	LSTM	Visual Genome, MS COCO	BLEU, METEOR, ROUGE-L, CIDEr, SPICE	56
241	J.S. Lu et al. 2018 [88]	ResNet, Faster R-CNN	LSTM	Flicker30k, MS COCO	BLEU, METEOR, CIDEr, SPICE	31
308	X.X. Lu et al. 2018 [15]	AlexNet,VGGNet, GoogLeNet	RNN, LSTM	UCM-captions, Sydney-captions, RSICD	BLEU, METEOR, ROUGE-L, CIDEr	17
320	Q. Wu et al. 2018 [89]	VGGNet	LSTM	Flicker8k/30k, MS COCO	BLEU, METEOR, CIDEr, sentence perplexity	15

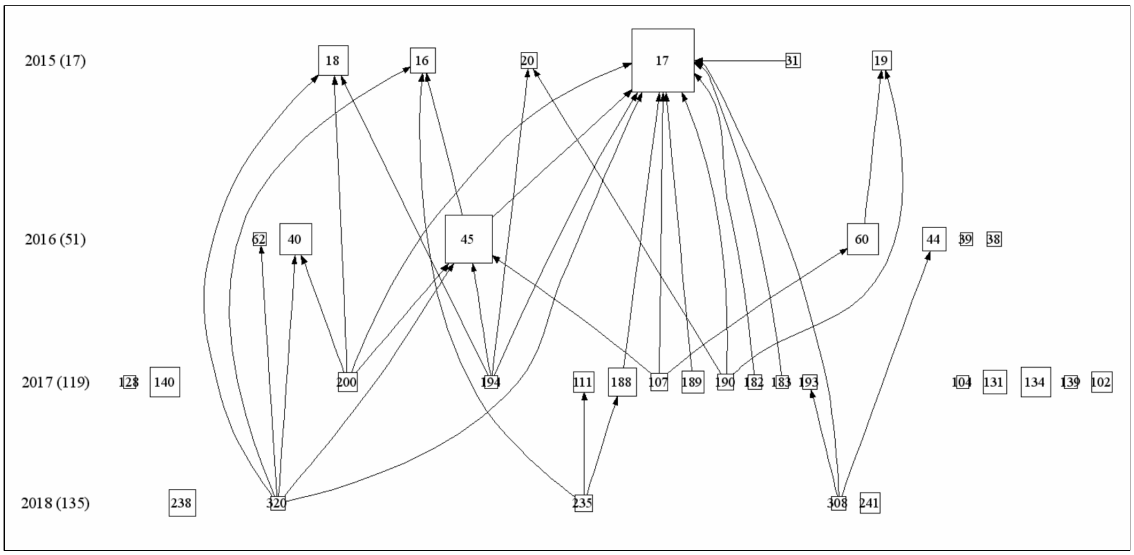


FIGURE 13. Timeline based map of top 35 references with 35 nodes and 34 links (The size of the nodes in the graph represents the number of times they have been cited, and the figures in the boxes represent the serial number of that document in HistCite software sorted by date).

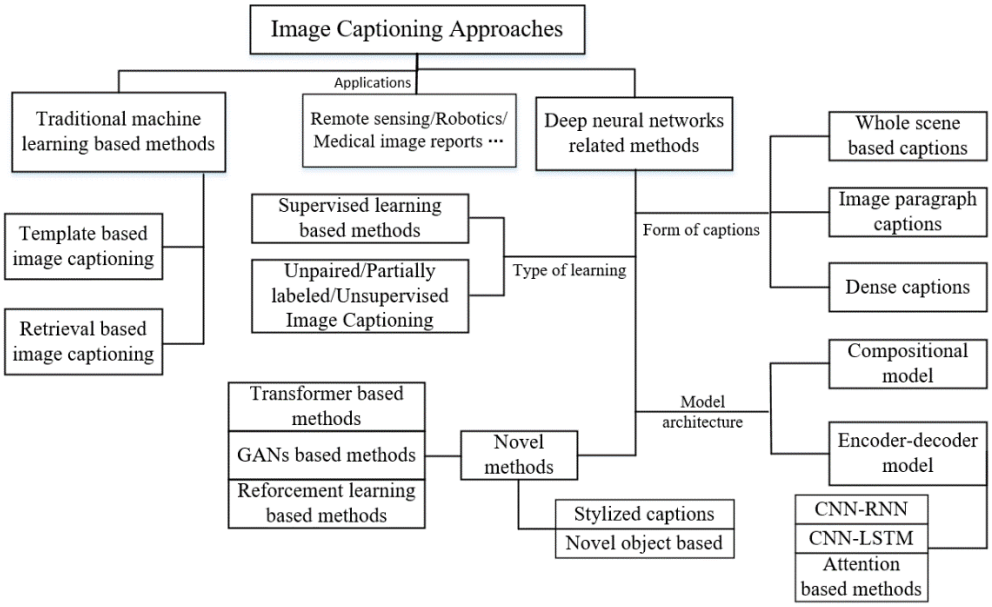


FIGURE 14. A taxonomy of image captioning approaches.

literature analysis and visualization results indicate that the output of publications in the field of image caption have increased year by year, while different attention mechanisms and Encoder-Decoder architectures have been the focus of development from 2015 to 2018.

A summary of image captioning models with their datasets and evaluation metrics are listed in Table 5. VGGNet and ResNet are the most commonly used visual models, and LSTM is the most popular linguistic model. As shown in the table, creating a richer and more applicable image caption dataset is a challenging task. Visual Genome [105] is a large-scale image semantic understanding dataset released in 2016, and it has become almost the standard dataset in the research of visual relationship detection. RSICD [15] is a large-scale benchmark data set of remote sensing, which can advance the task of remote sensing image captioning. Another challenging issue is the evaluation metrics. SPICE [36] is a novel semantic evaluation metric that measures how effectively image captions recover objects, attributes and the relations between them. Reference [79] proposed a new

metric SPIDEr, which is a linear combination of SPICE and CIDEr.

Therefore, combining the results of the previous studies, it could be seen that the evolutionary path of image captioning is:

(1) Prior to 2015, with the development of machine learning and the initial application of deep learning, image captioning was at a steady stage of development and it was not a hot topic in the field of vision research. In 2014, the Microsoft COCO was proposed, this dataset provided the basis for researchers to conduct research on image captioning.

(2) From 2015 to 2018, it was a period of rapid development for image captioning. Firstly, as part of the CVPR 2015 Large scale Scene Understanding workshop, the COCO Captioning Challenge is designed to spur the development of algorithms producing image captions. This captioning challenge has greatly boosted the research enthusiasm of scholars in the field. From the five highest cited articles in 2015, recurrent neural network visual representation model, neural image caption model, the gLSTM model, fast novel visual concept learning and video captioning are the hot spots of image captioning research at this stage. Secondly, after 2017, ResNet is used more often to train vision models, and the subdivision directions of image captioning models are gradually enriched, such as compositional captioning, dense captioning and attention based captioning. Among them, the studies based on different attention mechanisms are the main works in this phase [10], [13], [34], [50], [71], [75], [80], [83]–[85]. Thirdly, image captioning technologies have been successfully applied to remote sensing images, medical image reporting and robotics. And studies on datasets [15], [57] and evaluation methods [36] are also very important to drive the development of image captioning.

(3) From 2019 to now, image captioning is in a boom phase of development. New techniques such as Transformer, reinforcement learning and GANs have been widely applied to solve image description problems, and unsupervised image captioning methods [91]–[94] become a new research hotspot. The form of captioning has become more diverse as it is no longer confined to the overall content of the image [58], [81], [95]. In addition, Vision-Language Pretraining (VLP) model is an emerging direction of image captioning and image understanding. Reference [96] proposed a unified Vision-Language Pretraining model, which pre-trained on a large amount of image-text pairs using the unsupervised learning objectives of two tasks: bidirectional and sequence-to-sequence masked vision-language prediction. Reference [97] developed a new pre-trainable encoder-decoder structure that simultaneously supports both vision-language understanding and generation downstream tasks.

IV. DISCUSSION

Scientometric analysis studies are primarily designed to provide a broad under-standing of image captions because its methods rely on statistical analysis of bibliographic data.

Through the visualization of the scientometric analysis, we have recognized the spatial and temporal distribution characteristics, clarified the main research communities, understanding the current research hotspots and the scientific evolution paths of the image captioning field. For image description research, there is still a large gap between computer-generated text and human annotated ones, and automatic natural language generation will remain a challenging research topic for a long time [114].

How can bibliometric methods help the field of image captioning to develop a more scientific and better taxonomy? With the help of the previous analysis, we proposed a taxonomy of automatic image captioning methods. In Fig. 12, we group the different image caption approaches into two main categories, including traditional machine learning based image captioning and deep learning based image captioning. Then, we classify the existing deep neural network based methods into four categories based on ''type of learning'', ''form of captions'', ''model architecture'' and ''novel methods''. This systematic summary approach provides a more concise overview of the development of image captioning technology. Positive results have been achieved with current unsupervised and partially supervised image captioning techniques [91]–[94], [98] based on the question of whether training data is needed or whether image-text pairs are required. For the different forms of captions, in addition to the mainstream whole picture based caption, there are also image paragraph caption [99]–[103], and dense caption [8], [77], [104], [105]. From model architecture point of view, image captioning methods contain Encoder-Decoder architecture and compositional architecture [33], [76], [106]. Furthermore, novel image captioning methods are included, such as GANs based methods [81], [107], RL based methods [108], and Transformer based methods [109]–[113].

Every approach has limitations, and so does the method based on scientific measurement. Firstly, scientometric analysis may only be applied to those disciplines where literature and its citations are available from appropriate databases. As with many bibliometric methods, we have chosen only WOS to gain data. If there is insufficient published literature on a new research direction, it may not be possible to mine it through bibliometric methods. Secondly, although scientometric analysis is an empirical and objective method for analyzing knowledge structure, the interpretation of the graph is also important. There is also a need to understand the underlying algorithms and parameters in the different literature analysis tools so that the reader can read a good ''story''. For example, image captioning has only made great progress in the last few years, so the evolutionary path that can be shown on a macro level is relatively limited.

V. CONCLUSION

By analyzing bibliographic data in image captioning research, this article finds that the spatial and temporal distribution characteristics of image caption. The field of image description has shown a year-on-year increase in publications

over the last decade. China has published the largest amount of papers in this field, but the United States has had a greater impact on research in this area. Moreover, Microsoft Research, Google Research and other Silicon Valley giants, as well as universities such as the Australian National University and the Georgia Institution of Technology, have performed well in the field of image description.

In the meantime, we can answer the Q1-Q3 presented at the beginning of this article. (A1) Based on VOSviewer, we discover thirteen research communities. As a provider of the MSCOCO dataset, Microsoft researchers have produced numerous innovative results that form one of the key communities in the image caption field. Google Research presented the encoder-decoder architecture to researchers, and it is one of the most important communities in terms of impact. (A2) For hot topics and research trends mining, we find that model architecture, the evaluation metrics and datasets, new image caption models based on GANs, RL, Transformer, and remote sensing image captioning are the new research hotspots. (A3) So far, image captioning has gone through three major stages of development: (1) From 2010 to 2014, steady developing stage based on machine learning; (2) From 2015 to 2018, rapid maturation period based on deep neural networks; (3) From 2019 to now, prosperous phase of development based on new technologies.

This paper will support the scientific research in image captioning and help researchers to gain an understanding of the current state of the field, the active research community and research trends, so that they can promote the further development and use of image captioning technology.

REFERENCES

[1] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5561–5570.

[2] Y.-T. Chen, F. Chen, M. Cooper, and D. Joshi, "Using business-aware latent topics for image captioning in social media," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[3] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word-sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, early access, Dec. 25, 2020, doi: 10.1109/TGRS.2020.3044054.

[4] R. C. Luo, Y.-T. Hsu, Y.-C. Wen, and H.-J. Ye, "Visual image caption generation for service robotics and industrial applications," in *Proc. IEEE Int. Conf. Ind. Cyber Phys. Syst. (ICPS)*, May 2019, pp. 827–832.

[5] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 728–737.

[6] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional LSTMs," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 988–997.

[7] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4904–4912.

[8] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4904–4912.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, and A. Courville, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.

[11] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[12] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 203–212.

[13] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3242–3250.

[14] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.

[15] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

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

[17] X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning," *Vis. Comput.*, vol. 35, no. 3, pp. 445–470, Mar. 2019.

[18] A. Kumar and S. Goel, "A survey of evolution of image captioning techniques," *Int. J. Hybrid Intell. Syst.*, vol. 14, no. 3, pp. 123–139, Mar. 2018.

[19] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Feb. 2019.

[20] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, and E. Erdem, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, Jan. 2016.

[21] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018.

[22] A. Nasir, K. Shaukat, I. A. Hameed, S. Luo, T. M. Alam, and F. Iqbal, "A bibliometric analysis of corona pandemic in social sciences: A review of influential aspects and conceptual structure," *IEEE Access*, vol. 8, pp. 133377–133402, 2020.

[23] J. Niu, W. Tang, F. Xu, X. Zhou, and Y. Song, "Global research on artificial intelligence from 1990–2014: Spatially-explicit bibliometric analysis," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 5, p. 66, 2016.

[24] M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," *Social Sci. Inf.*, vol. 22, no. 2, pp. 191–235, 1983.

[25] G. Melin and O. Persson, "Studying research collaboration using co-authorships," *Scientometrics*, vol. 36, no. 3, pp. 363–377, Jul. 1996.

[26] K. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2389–2404, May 2010.

[27] E. Garfield, "Historiographic mapping of knowledge domains literature," *J. Inf. Sci.*, vol. 30, no. 2, pp. 119–145, 2004.

[28] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010.

[29] J. Li, F. Goerlandt, and G. Reniers, "An overview of scientometric mapping for the safety science community: Methods, tools, and framework," *Saf. Sci.*, vol. 134, Feb. 2021, Art. no. 105093.

[30] X. Y. An and Q. Q. Wu, "Co-word analysis of the trends in stem cells field based on subject heading weighting," *Scientometrics*, vol. 88, no. 1, pp. 133–144, Jul. 2011.

[31] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 359–377, 2006.

[32] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[33] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1141–1150.

[34] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[35] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 955–964.

[36] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 382–398.

[37] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1473–1482.

[38] J. Devlin, H. Cheng, H. Fang, S. Gupta, and L. Deng, "Language models for image captioning: The quirks and what works," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, 2015, pp. 100–105.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

[40] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," 2014, *arXiv:1410.1090*.

[41] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 11–20.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[43] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[44] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4534–4542.

[45] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "CapSal: Leveraging captioning to boost semantics for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6017–6026.

[46] H. Tan, F. Dernoncourt, Z. Lin, T. Bui, and M. Bansal, "Expressing visual relationships via language," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2019, pp. 1873–1883.

[47] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7378–7387.

[48] R. Luo, G. Shakhnarovich, S. Cohen, and B. Price, "Discriminability objective for training descriptive captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7378–7387.

[49] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, and R. Rossi, "Figure captioning with relation maps for reasoning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1526–1534.

[50] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2321–2334, Dec. 2017.

[51] K. Fu, J. Li, J. Jin, and C. Zhang, "Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5910–5921, Apr. 2018.

[52] X. Yu, T. Guo, K. Fu, L. Li, C. Zhang, and J. Zhang, "Image captioning with partially rewarded imitation learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[53] X. Dong, L. Zhu, D. Zhang, Y. Yang, and F. Wu, "Fast parameter adaptation for few-shot image captioning and visual question answering," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 54–62.

[54] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1029–1037.

[55] W. Zhao, W. Xu, M. Yang, J. Ye, Z. Zhao, Y. Feng, and Y. Qiao, "Dual learning for cross-domain image captioning," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 29–38.

[56] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1549–1557.

[57] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu, "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2347–2360, Sep. 2019.

[58] J. Aneja, H. Agrawal, D. Batra, and A. Schwing, "Sequential latent spaces for modeling the intention during diverse image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4260–4269.

[59] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12508–12518.

[60] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2407–2415.

[61] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5263–5271.

[62] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing novel objects in image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12489–12498.

[63] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[64] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, "Seeing bot," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 1341–1344.

[65] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2621–2629.

[66] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.

[67] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization (ACL)*, 2005, pp. 65–72.

[68] A. Agarwal and A. Lavie, "Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output," in *Proc. 3rd Workshop Stat. Mach. Transl. (ACL)*, 2008, pp. 115–118.

[69] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.

[70] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[71] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[72] X. R. Zhang, X. Wang, X. Tang, H. Y. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, p. 612, 2019.

[73] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2422–2431.

[74] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2533–2541.

[75] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.
[76] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.
[77] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.
[78] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 451–466.
[79] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 873–881.
[80] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1251–1259.
[81] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2989–2998.
[82] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.
[83] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.
[84] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4133–4139.
[85] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4176–4182.
[86] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
[87] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.
[88] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7219–7228.
[89] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.
[90] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell, "Understanding blind People's experiences with computer-generated captions of social media images," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 5988–5999.
[91] S. Cao, G. An, Z. Zheng, and Q. Ruan, "Interactions guided generative adversarial network for unsupervised image captioning," *Neurocomputing*, vol. 417, pp. 419–431, Dec. 2020.
[92] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4125–4134.
[93] J. Gao, Y. Zhou, P. L. H. Yu, S. Joty, and J. Gu, "Unsupervised cross-lingual image captioning," 2020, *arXiv:2010.01288*.
[94] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," 2020, *arXiv:2006.13611*.
[95] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1170–1178.
[96] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," 2019, *arXiv:1909.11059*.
[97] Y. Li, Y. Pan, T. Yao, J. Chen, and T. Mei, "Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network," 2021, *arXiv:2101.11562*.
[98] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7414–7424.
[99] X. He and X. Li, "Modeling coherence and diversity for image paragraph captioning," in *Proc. 5th Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Dec. 2020, pp. 634–639.
[100] R. Li, H. Liang, Y. Shi, F. Feng, and X. Wang, "Dual-CNN: A convolutional language decoder for paragraph image captioning," *Neurocomputing*, vol. 396, pp. 92–101, Jul. 2020.
[101] L. Melas-Kyriazi, A. M. Rush, and G. Han, "Training for diversity in image paragraph captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 757–761.
[102] Z. Wang, Y. Luo, Y. Li, Z. Huang, and H. Yin, "Look deeper see richer: Depth-aware image paragraph captioning," in *Proc. 26th ACM Int. Conf. Multimedia (MM)*, 2018, pp. 672–680.
[103] X. Yang, C. Gao, H. Zhang, and J. Cai, "Hierarchical scene graph encoder-decoder for image paragraph captioning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4181–4189.
[104] D.-J. Kim, T.-H. Oh, J. Choi, and I. So Kweon, "Dense relational image captioning via multi-task triple-stream networks," 2020, *arXiv:2010.03855*.
[105] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, and L.-J. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[106] R. Socher, Q. V. L. A. Karpathy, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, no. 1, pp. 207–218, 2014.
[107] H. Wang, Z. C. Qin, and T. Wan, "Text generation based on generative adversarial nets with latent variables," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, 2018, pp. 92–103.
[108] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1151–1159.
[109] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10575–10584.
[110] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," 2020, *arXiv:2004.14231*.
[111] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 11137–11147.
[112] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8927–8936.
[113] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, Dec. 2020.
[114] C. L. Chowdhary, A. Goyal, and B. K. Vasnani, "Experimental assessment of beam search algorithm for improvement in image caption generation," *J. Appl. Sci. Eng.*, vol. 22, no. 4, pp. 691–698, 2019.

WENXUAN LIU was born in 1987. He received the bachelor's degree in mathematics and applied mathematics from Northwest Normal University, Lanzhou, China, in 2010. He is currently pursuing the Ph.D. degree in cartography and geographical information engineering with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan, China. His current research interests include image captioning, remote sensing image understanding, and deep learning.

HUAYI WU was born in 1966. He received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1999. He worked as a Postdoctoral Fellow with the Geospace information and Communication Technology Laboratory, York University, Toronto, Canada, in 2002. He is currently a Professor and the Deputy Director of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan. His research interests include deep learning, data mining, and distributed computing. He is a member and the Secretary-General with the GIS Theory and Method Committee of China. He is also the Chairperson of the IV/2 working group of the International Society for Photogrammetry and Remote Sensing.

KAI HU was born in 1989. He received the Ph.D. degree in cartography and geographical information engineering from Wuhan University. He is currently working as an Associate Professor with the School of Internet of Things, Jiangnan University. His research interests include geographical information science, library information science, and scientometric. He has published ten articles in the related journals.

QING LUO was born in 1987. She received the Ph.D. degree in cartography and geographical information engineering from Wuhan University, Wuhan, China, in 2019. She currently works as a Lecturer with the Department of Big Data Science, School of Mathematics and Physics, Wuhan Institute of Technology. Her research interests include spatial statistics, spatial analysis, and spatial big data analysis.

XIAOQIANG CHENG was born in 1985. He received the Ph.D. degree from Wuhan University. He is currently working as a Lecturer with the Faculty of Resources and Environmental Science, Hubei University. His research interests include geovisualization, cartography, and geospatial web.