Deep Learning Approaches on Image Captioning: A Review

TARANEH GHANDI, McMaster University, Canada
HAMIDREZA POURREZA, Ferdowsi University of Mashhad, Iran
HAMIDREZA MAHYAR, McMaster University, Canada

Image captioning is a research area of immense importance, aiming to generate natural language descriptions for visual content in the form of still images. The advent of deep learning and more recently vision-language pre-training techniques has revolutionized the field, leading to more sophisticated methods and improved performance. In this survey article, we provide a structured review of deep learning methods in image captioning by presenting a comprehensive taxonomy and discussing each method category in detail. Additionally, we examine the datasets commonly employed in image captioning research, as well as the evaluation metrics used to assess the performance of different captioning models. We address the challenges faced in this field by emphasizing issues such as object hallucination, missing context, illumination conditions, contextual understanding, and referring expressions. We rank different deep learning methods' performance according to widely used evaluation metrics, giving insight into the current state-of-the-art. Furthermore, we identify several potential future directions for research in this area, which include tackling the information misalignment problem between image and text modalities, mitigating dataset bias, incorporating vision-language pre-training methods to enhance caption generation, and developing improved evaluation tools to accurately measure the quality of image captions.

CCS Concepts: • **Computing methodologies** → **Supervised learning**; **Unsupervised learning**; **Reinforcement learning**; *Neural networks*; **Scene understanding**; **Natural language generation**; Machine translation;

Additional Key Words and Phrases: Image captioning, deep learning, text generation

ACM Reference format:
Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep Learning Approaches on Image Captioning: A Review. *ACM Comput. Surv.* 56, 3, Article 62 (October 2023), 39 pages.
https://doi.org/10.1145/3617592

1 INTRODUCTION

Automatic image captioning is a critical research problem with numerous complexities, attracting a significant amount of work with extensive applications across various domains such as human-computer interaction [32, 71, 143], medical image captioning and prescription [9, 58, 96], traffic data analysis [69], quality control in industry [83], and especially assistive technologies for visually impaired individuals [2, 27, 46, 85, 110]. The field has undergone a revolutionary

Authors' addresses: T. Ghandi and H. Mahyar (Corresponding author), McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4L8, Canada; e-mails: {ghandit, mahyarh}@mcmaster.ca; H. Pourreza, Ferdowsi University of Mashhad, Azadi Square, Mashhad, Razavi Khorasan Province, Iran, 9177948974; e-mail: hpourreza@um.ac.ir.

0360-0300/2023/10-ART62 $15.00
https://doi.org/10.1145/3617592

transformation with the development and growth of deep learning techniques [3, 4, 99, 115, 116], resulting in the emergence of advanced methods and enhanced performance. Automatic image captioning lies at the intersection of natural language processing and computer vision. This field of research deals with the creation of textual descriptions for images without human intervention. Given an input image I, the goal is to generate a caption C describing the visual contents present inside the given image, with C being a set of sentences $C = \{c_1, c_2, \ldots, c_n\}$ where each c_i is a sentence of the generated caption C.

Given the recent advancements in the domain of image captioning, an updated review of the more recent research works can assist researchers in keeping up with the latest progress in this field. There exist numerous literature reviews and surveys on image captioning, providing an extensive collection of research conducted in previous years. Notably, Hossain et al. [52] authored a comprehensive survey paper that served as an inspiration for this work's structure. However, instead of a pairwise comparison like in Reference [52], we have organized our article to feature a separate section for each method category and follow the same order of category in our discussion (Section 4). Furthermore, most surveys typically cover works dating from 2018 and earlier, while more recent research is yet to be addressed. Some surveys [29] are limited in the number of research works covered, while others [20] do not delve into the methodologies' specific details. Additionally, considering the recent advancements of vision language pre-training methods, image captioning methods that fall under this category must be addressed and discussed. This category has seldom been explored in previous survey works. The field of image captioning can be classified into multiple categories that differ in the captioning settings, such as dense captioning methods that provide captions for each entity presented in the image or whole image captioning methods that provide captions for the entirety of the input image. Here, we focus on reviewing "whole image" captioning methods.

In this article, we discuss various methods of image captioning introduced in papers published from 2018 to 2022, followed by the most common problems and challenges of image captioning. We provide a comprehensive analysis of each method, covering widely used datasets and evaluation metrics. We also compare the performance of the different covered methods before exploring future directions in the field. The section on problems and challenges provides a detailed overview of the inherent difficulties in image captioning and provides insight into potential solutions to address them. We hope to provide a thorough understanding of image captioning through this review and encourage continued progress in the field.

2 DEEP LEARNING-BASED IMAGE CAPTIONING

In this section, we have organized and classified the different frameworks, methods, and approaches that were extensively used in recent research works based on their core structure. Some terms and notations in the covered papers have been altered to maintain consistency throughout this review. A figure demonstrating the taxonomy provided in this article is shown in Figure 1.

2.1 Attention-based Methods

The methods that fall under the attention-based category utilize attention mechanisms to emphasize the most relevant parts of the input image when generating captions.

Attention-based methods [10] are inspired by the human attention pattern and the way the human eye focuses on images. When inspecting images, humans focus more on the image's salient features. The same mechanism is implemented in attention-based mechanisms. During the training process, the model is shown "where to look at." To understand the mechanism of attention-based methods, one can imagine a sequential decoder in which, in addition to the previous cell's output and internal state, there is also a context vector under the term "c."

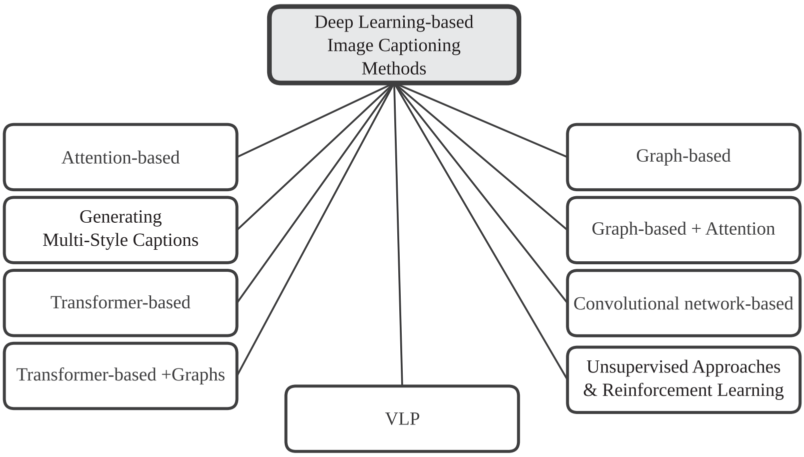

Deep Learning-based Image Captioning Methods
Attention-based
Generating Multi-Style Captions
Transformer-based
Transformer-based +Graphs
VLP
Graph-based
Graph-based + Attention
Convolutional network-based
Unsupervised Approaches & Reinforcement Learning

Fig. 1. The taxonomy of the image captioning methods covered in this survey article.

Vector c is the weighted sum of hidden states in the encoder.

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{1}$$

In the statement above, a_{ij} is the "amount of attention" that output i must pay to input j, and h_j is the encoder state in input j. a_{ij} is obtained by calculating softmax over attention amounts that are shown with e on inputs and for output i:

$$a_{ij} = \text{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \tag{2}$$

$$e_{ij} = f(S_{i-1}, h_j), \tag{3}$$

where f is the model that determines how much input at j and output i are correlated, and S_{i-1} is the hidden state from the previous timestep. The model f can be estimated with a small neural network and can be optimized with any gradient-based optimization techniques, such as gradient descent.

In short, attention-based image captioning methods generate a weighted sum of extracted feature vectors at each timestep in their decoder that guides the decoder module. Similar to the encoder-decoder framework, attention-based methods were first introduced for the machine translation problem in Reference [10]. In most of the attention-based methods, a CNN or a region-based CNN is used in the encoding stage to provide a representation of the image, and an RNN is usually used in the decoding stage. A block diagram of the basis of attention-based methods (which was first proposed by Xu et al. [128]) is shown in Figure 2. The last layer of a Convolutional Neural Network (here, VGGnet by Simonyan et al. [111])—just before max pooling—has been used to extract features from the image. The LSTM network [51] with attention has been used as the decoder. The multiple images surrounding the LSTM shown in this figure demonstrate the attention values over different regions of the image. The lighter areas mean a higher attention value. The colored outline of the generated words in the caption corresponds to the regions outlined by the same colors.

Attention-based methods are widely used in the encoder-decoder framework. Most of the research works discussed in this survey have used it as their primary framework or have combined it

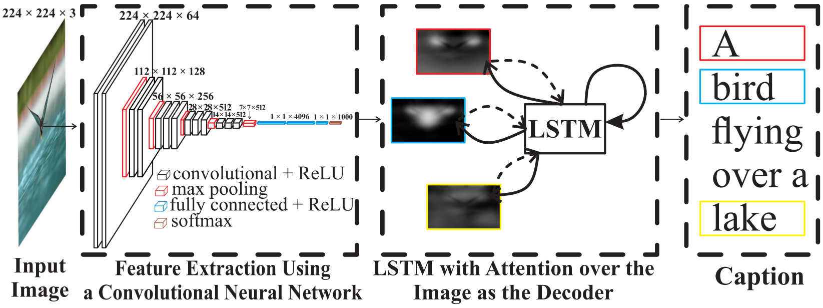


Fig. 2. The basis of attention-based methods (best viewed in color).

with other methods to improve its performance. Vinyals et al. [119] have been the first to incorporate deep learning-based encoder-decoder framework for image captioning. The presented model in their work is inspired by machine translation, based on the findings that indicate that given a powerful sequence model, it is possible to achieve remarkable results by directly maximizing the probability of the correct translation. CNNs can produce a rich presentation of an input image by embedding it into a fixed-length vector. Vinyals et al. [119] have presented a model that uses CNN as an image “encoder” by pre-training it for an image classification task first and using the last hidden layer as an input to an RNN “decoder” that generates sentences. The model is trained to maximize the likelihood of the target description sentence given the training image. This work has been used by many other researchers as a basis to expand upon and refine using other modules and techniques [128].

Anderson et al. proposed the “bottom-up and top-down” method in Reference [7]. The bottom-up module proposes the salient regions in the image, and each of the proposed regions is represented as a convolutional feature vector. This module is implemented using Faster R-CNN [102], which was discussed previously. Faster R-CNN works well as a “hard” attention mechanism, since a small number of bounding box features are selected from a large number of configurations. Faster R-CNN network is initialized with ResNet-101 [47] pre-trained for image classification on the ImageNet dataset. Faster R-CNN is then trained using the Visual Genome [63] dataset. The top-down module, designed to caption images, contains two LSTM networks [51] with the standard implementation. The first LSTM network operates as a top-down visual attention model, and the second LSTM network operates as a language model. The top-down visual attention module estimates a distribution of attention over regions and calculates the extracted feature vector as a weighted sum over total region proposals. The captioning model takes a variably sized set of k image features: $V = \{v_1, \ldots, v_k\}, v_i \in \mathbb{R}^D$ as input. Each image feature encodes a salient region of the image. These image features can be defined as the output of the bottom-up attention model or as the spatial output layer of a CNN. The input vector to the attention LSTM at each timestep consists of the previous output of the language LSTM, the mean-pooled image features $\bar{v} = \frac{1}{k} \sum_i v_i$, and an encoding of the word generated previously all concatenated together. The input to the language model LSTM is composed of the attended image feature concatenated with the output of the attention LSTM.

The two-layer LSTM [51] structure has also been used by Yao et al. in Reference [136] as the attention mechanism in the final stage. (More detail on the workings of this paper is discussed in “Combining Attention-Based and Graph-Based Methods” (Section 2.3).)

Gu et al. [41] have presented a multi-stage coarse-to-fine structure for image captioning. This structure contains multiple decoders that each work on the output of the decoder in the previous step, making the captions richer in every step. This paper has used the LSTM network [51] as the decoder. The structure comprises three LSTM networks, with the first LSTM presenting the coarse details at the first stage and reducing computations in the later stages. The other LSTMs operate as fine-level decoders. At each stage, attention weights and hidden vectors generated by the previous stage decoder are used as input to the next stage decoder.

The operation of the coarse decoder is based on the general and global features of the image. However, in many cases, each word belongs to a small region of the image only. Using the general features of the image might yield improper results due to the possible noise from unrelated regions. Therefore, a "Stacked Attention Model [135]" is used to improve the performance of this coarse-to-fine structure. This model enables the structure to extract visual information from finer details for future word predictions. The stacked model generates a spatial map that determines the region of each predicted word. Using this stacked attention model, finer and more precise details are extracted, and noise is gradually reduced. Also, regions that are highly relevant to the words are determined.

Huang et al. [59] have introduced a new attention-based structure containing one more level of attention. The structure, named **Attention on Attention (AoA)**, generates an "Information Vector" and an "Attention Gate" with two linear transformations. An attention module $f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ operates on some queries, keys, and values denoted by $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively, and generates some weighted average vectors denoted by $\hat{\boldsymbol{V}}$. The attention module measures the similarity between $\boldsymbol{Q}$ and $\boldsymbol{K}$ and uses this similarity score to calculate weighted average vectors over $\boldsymbol{V}$, which is formulated as:

$$a_{i,j} = f_{sim}(\boldsymbol{q_i}, \boldsymbol{k_j}), \alpha_{i,j} = \frac{e^{a_{i,j}}}{\sum_j e^{a_{i,j}}}, \tag{4}$$

$$\hat{\boldsymbol{v_i}} = \sum_j \alpha_{i,j} \boldsymbol{v_j}, \tag{5}$$

where $\boldsymbol{q_i} \in \boldsymbol{Q}$ is the ith query, $\boldsymbol{k_j} \in \boldsymbol{K}$ and $\boldsymbol{v_j} \in \boldsymbol{V}$ are the jth key/value pair. f_{sim} is a function that computes the similarity score of each $\boldsymbol{k_j}$ and $\boldsymbol{q_i}$, and $\hat{\boldsymbol{v_i}}$ is the attended vector for the query $\boldsymbol{q_i}$. Since the attention module produces a weighted average for each query regardless of the relation between $\boldsymbol{Q}$ and $\boldsymbol{K}/\boldsymbol{V}$, the weighted average vector can be irrelevant or misleading information. The AoA module measures the relevance between the attention results and the query. The information vector $\boldsymbol{i}$ is generated with a linear transformation on current content (caption) and results from the attention component and stores both parts' data. The attention gate $\boldsymbol{g}$ is generated from the content and the result from the attention component using sigmoid activation. The value inside each part (also called a channel) of this attention gate determines the level of importance of the channel in the information vector. Both the information vector and the attention gate are conditioned on the attention result and the current context (i.e., the query) $\boldsymbol{q}$. The AoA structure adds another level of attention with element-wise multiplication of the attention gate and the information vector and finally produces the attended information, which contains useful data. The AoA structure is applied to both the encoder and decoder (termed AoANet): AoA is applied to the encoder after extracting image features to obtain the relation between objects present inside the image. AoA is also applied to the decoder to remove the attention results that are unrelated to the actual output or are ambiguous and leave the essential and useful results. The AoA structure has been introduced as an addition to the attention-based methods, and it can be applied to any attention method. In the experiments conducted by the authors, a Faster R-CNN [102] pre-trained on the ImageNet [25] and Visual Genome [63] datasets is used to extract feature vectors from the image.

Jiang et al. propose a novel **recurrent fusion network (RFNet)** in Reference [61] for the image captioning task, which uses multiple CNNs as encoders, and a recurrent fusion process is inserted after the encoders to produce better representations for the decoder. Each representation extracted from an individual image CNN can be regarded as an individual view depicting the input image. The fusion procedure consists of two stages: The first stage produces multiple sets of "thought vectors" by exploiting the interactions among the representations from multiple CNNs; the second stage performs multi-attention on the sets of thought vectors and generates a new set of thought vectors for the decoder. For the experiments, they use ResNet [47], DenseNet [57], Inception-V3 [113], Inception-V4 [112], and Inception-ResNet-V2 [112] as encoders to extract five groups of representations. Having considered **reinforcement learning (RL)** as a method to improve image captioning performance, they have trained their model with cross-entropy loss and fine-tuned the trained model with CIDEr optimization using reinforcement learning.

Incorporating attention in image captioning has transformed the field considerably, enabling more accurate and natural caption generation. However, they do not come without flaws. One problem with classic attention-based image captioning is that they do not consider the relations between the objects detected inside the image.

2.1.1 Injecting Spatial and Semantic Relation Information into Attention-based Methods. A group of attention-based methods has included the spatial and semantic relations in an image to describe the content more appropriately.

Pan et al. [93] introduced a novel attention method termed the "X-Linear Attention Block," which emphasizes salient image features and supports multimodal reasoning through the use of bilinear pooling. This structure employs spatial and channel-wise bilinear attention to extract second-order interactions. These interactions are computed by taking the outer product between the key (representing mapped image features) and the query (representing the internal state of the sentence decoder) using bilinear pooling to capture all second-order interactions. Following bilinear pooling, two embedding layers predict attention weights for each region, which are then normalized using a softmax layer to obtain the spatial attention vector. A "squeeze-excitation" operation is performed on the embedded outer product (feature map). The squeezing process aggregates the feature map across spatial regions, generating a channel descriptor. The excitation process employs a self-gating mechanism with a sigmoid function on the channel descriptor, resulting in the channel-wise attention vector. Finally, the outer product of the key and query, along with the value from bilinear pooling, is weighted summated with the spatial attention vector. The resulting weighted sum undergoes channel-wise multiplication with the channel attention vector, yielding the attended features. Higher-order interactions can be computed by combining multiple X-Linear attention blocks. In this work, Faster R-CNN [102] is employed for region detection. A stack of X-Linear attention blocks is then utilized to encode the region-level features of the image and capture higher-order interactions between these regions. This process generates a set of enhanced region-level and image-level features. The attention blocks are further integrated into the sentence decoder to facilitate multimodal reasoning.

Cornia et al. [22] presented a method capable of describing an image by focusing on different regions in different orders following a given conditioning. By means of analyzing the syntactic dependencies between words, a higher level of abstraction can be recovered in which words can be organized into a tree-like structure. In a dependency tree, each word is linked together with its modifiers. Given a dependency tree, nouns can be grouped with their modifiers, thus building *noun chunks*. The proposed model is built on a recurrent architecture that considers the decomposition of a sentence into noun chunks and models the relationship between image regions and textual chunks to ground the generation process on image regions explicitly. The model is conditioned on the input image I and an ordered sequence of region sets R, which acts as a control signal and jointly predicts two output distributions corresponding to the word-level and chunk-level representation of the sentence. During the generation, the model keeps a pointer to the current region and can shift to the next element in R using a Boolean chunk-shifting gate g_t. To generate the output caption, a recurrent neural network with adaptive attention is used. The probability of switching to another chunk $p(g_t|R)$ is calculated in an adaptive mechanism in which an LSTM [51] computes a compatibility function between its internal state and a latent representation modeling the state of memory at the end of a chunk. The compatibility score is compared to that of attending one of the regions r_t, and the result is used as an indicator to switch to the next region set in R.

The addition of spatial and semantic relations to the attention-based framework has significantly improved the quality of the captions generated by the models. Despite the improvements achieved by this addition, some problems still remain, including the ambiguity of the captions, the lack of grounding, heavy computations associated with the object detectors, and the requirement of bounding-box annotations. To resolve some of these issues, other approaches to the image captioning problem have been introduced, which are explained and discussed in the following sections.

2.2 Graph-based Methods for Spatial and Semantic Relations between Image Elements

The methods discussed in this section utilize scene graphs to better model the spatial and semantic relations between image elements.

Due to their ability to represent relations between elements, graphs are used in applications in which the relations between elements are important [13, 120]. Studies have shown the effectiveness of incorporating semantic information and object attributes in generating captions of higher quality [34, 125, 137, 138, 147]. Some research works on image captioning have used graphs to incorporate the spatial and semantic relations between the elements inside an image. To utilize graphs in caption generation, two types of graph extraction are usually used: scene graph extraction from images [24, 44, 73, 114, 127, 132, 142] and scene graph extraction from textual data [6, 124]. Once a scene is abstracted into symbols, the language generation is almost independent of visual perception [134]. Given scene abstractions "helmet-on-human" and "road dirty," humans can infer "a man with a helmet in the countryside" by using common sense knowledge like "countryside road dirty." This can be considered as the inductive bias that enables humans to perform better than machines.

Yang et al. [134] have integrated the inductive bias of language generation into the encoder-decoder framework commonly used in image captioning. The proposed method uses scene graphs to connect the image and text modalities. A scene graph G is a unified representation that connects the objects, their attributes, and their relationships in an image I or a sentence S by directed edges. To encode the language prior, Yang et al. [134] proposed the **Scene Graph Auto-Encoder (SGAE)**, which is a sentence self-reconstruction network used in the $I \rightarrow G \rightarrow D \rightarrow S$ training pipeline. The $I \rightarrow G$ module is a visual scene graph detector. A multi-modal GCN is introduced and used in the $G \rightarrow D$ module to complement the visual cues that may be ignored due to imperfect visual detection. D can be considered as a working memory [118] that assists in re-keying the encoded nodes from I to S to a more generic representation with smaller domain gaps. The proposed SGAE-based image captioning model is implemented using Faster R-CNN [102], and the language decoder proposed by Reference [7] with RL-based training strategy [103]. The proposed framework is formulated as follows:

$$\begin{gathered}\boldsymbol{Encoder}: V \leftarrow I, \\ \boldsymbol{Map}: \hat{V} \leftarrow R(V, G; D), G \leftarrow V, \\ \boldsymbol{Decoder}: S \leftarrow \hat{V},\end{gathered} \tag{6}$$

where V denotes the extracted image features (usually extracted by a **Convolutional Neural Network (CNN)**). The mapping module frequently used in the encoder-decoder framework for

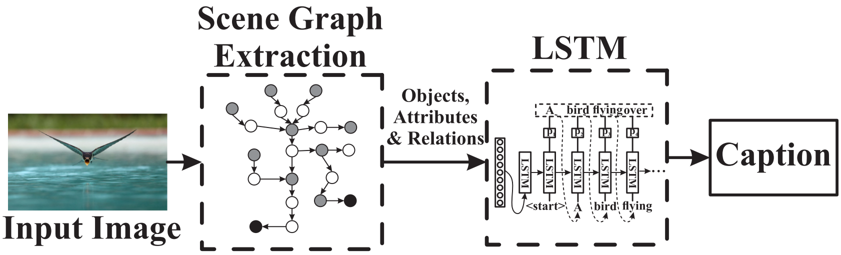


Fig. 3. The general workflow of graph-based methods.

image captioning is the module that encodes the visual features from the image into a representation that is later taken as input by the language decoder. This mapping module has been modified according to the formulation in Equation (6) by introducing the scene graph G into a re-encoder R parametrized by a shared dictionary D. The **Scene Graph Auto-Encoder (SGAE)** learns the dictionary D, which embeds the language inductive bias from sentence-to-sentence reconstruction. Next, the encoder-decoder framework is equipped with SGAE to form the overall image captioner.

Gu et al. [43] have introduced a particular framework for training an image captioning model in an unsupervised manner and without image-caption pairs. The framework uses a scene graph to generate an intermediate representation of images and captions and maps these scene graphs to their feature space using "Cycle-Consistent Adversarial Training" [148]. This paper has used an image scene graph generator, a sentence scene graph generator, and a feature mapping module in charge of mapping image features and captions modalities together. To align scene graphs and captions, CycleGAN [148] is used. The unrelated image and sentence scene graphs are first encoded using the scene graph encoder trained on the sentence corpus. Next, unsupervised cross-modal mapping is performed for feature alignment with CycleGAN. This work is closely related to Reference [134]. The main difference is that the framework in Reference [134] is based on paired settings. CycleGAN is generally used to transform two images together, and one of its applications is transforming two image elements together (for example, transforming an apple into an orange or a horse into a zebra).

Gao et al. [35] proposed a scene-graph-based semantic representation method by embedding the scene graph as an intermediate state. The task of image captioning is divided into two phases termed: concept cognition and sentence reconstruction. In the first phase, a vocabulary of semantic concepts is built, and a novel CNN-RNN-SVM framework is used to generate a scene-graph-based semantic representation, which is used as the input for an RNN generating captions in the second phase. The CNN part extracts visual features, the RNN part models image/concept relationships and concept dependency, and the SVM part classifies the semantic concepts and outputs the relevant concepts for the scene-graph-based sequence.

The general workflow of the graph-based methods is displayed in Figure 3. Usually, a Convolutional Neural Network is used to extract visual features from the image, and the semantic and spatial graph is built on the detected regions. The vertices denote regions, and the edges denote the relationships between the regions. Next, **Graph Convolutional Networks (GCNs)** [62] encode the regions and relationships in the scene graph. The obtained feature vector is then passed to LSTM [51] decoders to generate captions.

2.3 Combining Attention-based Methods and Graph-based Methods

To solve some of the issues revolving around image captioning problems and the problems regarding attention-based and graph-based methods, some recent research works have introduced structures that combine the two methodologies.

As previously mentioned, the visual relations between image elements give insight into their relative positions or interactions. To detect the visual relations between image elements, one not only needs to detect object locations inside the image, but also needs to detect all sorts of interaction between pairs of elements. Using these visual relations will allow for a more in-depth comprehension of images. However, the considerable diversity in object sizes and their locations will make the interaction detection task more difficult.

Yao et al. [136] use a combination of Graph Convolutional Networks [62] and LSTMs [51] to incorporate the relations between image elements while also taking the attention-based encoder-decoder framework into account. Spatial and semantic relations have been integrated to enrich image representations in the image encoder, and learning the relationships has been considered a classification problem. Faster R-CNN [102] has been used for region proposals. Two spatial and semantic graphs are built to represent spatial and semantic relations between image contents. These two graphs are generated from the detected image regions, with regions being graph nodes and the relations between them as the edges of the graph. In the spatial graph, spatial relations are considered as edges, and in the semantic graph, the semantic relations are considered as edges. The semantic graph is trained using the Visual Genome [63] dataset. To represent the image, Graph Convolutional Networks [62] are used that incorporate the semantic and spatial relations obtained from their corresponding graphs. The combination of the enhanced image region representations and their semantic and spatial relations are then fed into an LSTM [51] decoder to generate the caption sentences. During inference, to combine the output of the two spatial and semantic decoders, the distribution over the words generated by the two decoders is linear weight summated at each timestep, and the word with the highest probability is extracted.

The proposed model by Zhong et al. [145] decomposes the image scene graph into a set of sub-graphs. Each sub-graph captures a semantic component of the input image. Zhong et al. [145] designed a **sub-graph proposal network (sGPN)** that learns to detect meaningful sub-graphs. An attention-based LSTM then decodes the selected sub-graphs for generating sentences. Given an input image I, a scene graph $G = (V, E)$ is extracted from I using MotifNet [140], where V represents the nodes corresponding to the detected objects in I and E represents the set of edges corresponding to the relationships between object pairs. The goal is to generate a set of sentences $S = \{S_j\}$ to describe different components of the image using the scene graph G. Sub-graphs are defined as $\{G_i^c = (V_i^c, E_i^c)\}$, where $V_i^c \subseteq V$ and $E_i^c \subseteq E$. The method aims to model the joint probability $P(C_{ij} = (G, G_i^c, C_j)|I)$, where $P(C_{ij}|I) = 1$ when the sub-graph G_i^c can be used to decode the sentence S_j and $P(C_{ij}|I) = 0$ otherwise. $P(C_{ij}|I)$ can be decomposed into three parts:

$$P(C_{ij}|I) = P(G|I)P(G_i^c|G, I)P(S_j|G_i^c, G, I). \tag{7}$$

$P(G|I)$ can be interpreted as the scene graph extraction phase, $P(G_i^c|G, I)$ as the scene graph decomposition phase and the selection of important sub-graphs for sentence generation, and $P(S_j|G_i^c, G, I)$ as the decoding phase in which a selected sub-graph G_i^c is decoded into its corresponding sentence S_j, and the tokens in S_j are associated to the nodes V_i^c of the sub-graph G_i^c (the image regions in I).

Wang et al. [121] have used a Graph Neural Network [105] to represent the relation between image elements and have used a novel content-based attention framework to store image regions previously attended by the attention module as well. A ResNet-101 [47] network trained on the ImageNet [25] dataset is first used to extract image features. The non-linear activations of the last convolutional layer of this network are used as the image representation and are denoted as:

$$V = \{v_1, v_2, \ldots, v_n \mid v_i \in \mathbb{R}^m\}, \tag{8}$$

where v_i represents each of the non-linear activations of the last convolutional layer. A Graph Convolutional Network [62] f_{gnn} is initialized using the image features belonging to different image regions to explore relations between the visual objects in the image. This graph initializes each node inside the graph with a spatial representation and to derive the implicit relation-aware representation $R = \{r_1, r_2, \ldots, r_n \mid r_i \in \mathbb{R}^m\}$ (where r_i represents the nodes inside the graph), updates the value of the nodes with hidden representation from other nodes recursively. The visual representations R are forwarded into context-aware attention model f_{att}. Unlike some other attention-based models, this novel attention framework uses LSTM [51] to store the previously attended regions. Storing these regions will aid the attention module in its future region selections. Next, a language model based on LSTM, f_{lstm}, uses the previous hidden state $h_{(t-1)}$, the previously generated word embedding X_t, and the output $\bar{v}_t$ from the attention model as input and produces the current hidden state h_t as the output to predict the next word.

Chen et al. [18] have proposed a model to generate controllable image captions that actively consider user intentions. The paper introduces a more fine-grained control signal called **Abstract Scene Graph (ASG)**, a directed graph composed of three types of abstract nodes grounded in the image: object, attribute, and relationship. The caption generation model is based on the encoder-decoder framework, consisting of a role-aware graph encoder and a language decoder that considers both the context and structure of nodes for attention. The decoder utilizes a two-layer LSTM [51] structure, including an attention LSTM and a language LSTM. The model gradually updates the graph representation during decoding to fully cover information in ASG without omission or repetition and keep track of graph access status. The role-aware graph encoder contains a role-aware node embedding to distinguish node intentions and a multi-relational Graph Convolutional Network for contextual encoding.

Aiming to employ knowledge in scene graphs for image captioning explicitly, Li et al. [70] introduce a framework based on scene graphs. First, the scene graph for the input image is generated using the method proposed in Reference [127]. A set of initial bounding boxes should be produced to generate the scene graph. Li et al. have used the **region proposal network (RPN)** proposed by Girshick et al. [38] to produce a set of object proposals for the image. To capture the visual features, the VGG-16 network is used to extract CNN features from the corresponding regions of object entities. Semantic features are also obtained by extracting triplets, which are lexeme sequences that describe object relationships from the graph and embed them into fixed-length vectors. To utilize both types of information, a hierarchical attention-based fusion module is introduced that determines when and what to attend to during sentence generation.

Xu et al. [129] proposed a framework to embed the scene graph into a compact representation capable of capturing explicit semantic concepts and graph topology information. An input image I is processed by a CNN to generate the image features. A set of modules detect the objects, attributes, and related components to infer the scene graph. Next, an external vocabulary compiles the scene graph into the vector $V_{con}(I)$. An adjacent matrix is presented where the objects and relationships of the graphs are used as vertices and edges. A fixed-length vector $V_{topo}(I)$ is extracted to capture the topological information from the adjacent matrix. Xu et al. proposed an attention extraction mechanism that extracts sub-graphs and selects an attention graph with the corresponding region by computing cluster nodes in the adjacency matrix. The attention region is denoted as $V_{att}(I)$. The four vectors are combined into a single representation for the scene graph, which is fed into the LSTM-based [51] language model.

Lee et al. [65] have extended the top-down captioner introduced in Reference [7] and have added an attention component for relation features. No graph convolutions are used in the proposed model. The authors state that using visual relations from scene graphs directly is an alternative to GCNs and avoids expensive graph convolutions.

There is a different set of challenges associated with the use of scene-graphs. Scene graph extraction is a difficult task on its own, and the relations between the elements are not always as simple as pairwise relationships. Graph parsers still need improvement as well.

2.4 Convolutional Network-based Methods

Convolutional network-based methods utilize convolutional neural networks to extract image features and generate captions using output from a language model. Thanks to the recent advances in convolutional architectures on other sequence-to-sequence tasks such as convolutional image generation [90] and machine translation [36, 37], it is possible to consider CNNs as an effective solution to many vision-language tasks. The methods discussed in this section have incorporated CNNs into their proposed systems.

Inspired by the advances of CNNs in vision-language tasks, Aneja et al. [8] have presented a convolutional model containing three main components. The first and the last components are input/output word embeddings, respectively. While the middle component contains LSTM [51] or GRU [19] units in other methods, masked convolutions are used in the proposed approach. This component is feed-forward without any recurrent functions, unlike the RNN approaches.

Wang et al. [122] proposed a framework relying on Convolutional Neural Networks only to generate captions. The framework consists of four modules: a vision module, a language module, an attention module, and a prediction module. The vision module is a CNN without the fully connected layer, for which VGG-16 has been used. The language module is based on a CNN without pooling. RNNs use a recurrent path to memorize context, whereas CNNs use kernels and stack multiple layers to model the context. The prediction module is a one-hidden layer neural network as well. Since different levels of the language CNN represent different levels of concept, a hierarchical attention module has been employed where attention vectors are calculated at each level of the language model and fed into the next level. Since the attention maps are computed in a bottom-up manner as opposed to the RNN-based model, it is possible to train the model in parallel over all words in the sentence. The authors observed the effect of several hyper-parameters, such as the number of layers and the width of the kernel belonging to the language CNN. The receptive field of the language CNN can be increased by stacking more layers or increasing the width of the kernel. The experiments showed that increasing the kernel width is a better choice.

Less attention has been paid to convolutional network-based methods compared to the categories discussed above. Convolutional network-based models help generate more entropy and, as a result, more caption diversity. Also, they perform better in classification tasks and do not suffer from vanishing gradients. However, these methods still need improvement in terms of performance according to the evaluation metrics.

2.5 Transformer-based Methods

Many current works have utilized Transformers to build more robust solutions for the captioning problem. RNNs and LSTMs have been criticized due to their inflexibility, limitations regarding expression ability, and other complexities. Due to their recurrent nature, RNNs have difficulty memorizing inputs many steps ago, which leads to high-frequency phrase fragments without regard to the visual cues [67]. The limitations posed by LSTMs and RNNs as language models have led researchers to use alternatives such as Transformers.

Some recent works have studied the application of Transformers [116]—mainly as the language model. Herdade et al. [49] utilized Transformers in the proposed "Object Relation Transformer" model, which incorporates spatial relations between detected objects using geometric attention. This encoder-decoder-based structure implements spatial relationships between detected objects inside an image using geometric attention. The object relation module presented by Hu et al. [53] represents the spatial relations in the encoder. The combination of Faster R-CNN [102], and ResNet-101 [47] as the base Convolutional Neural Network is used for object detection and feature extraction. Every image feature vector is processed through an input embedding layer consisting of a fully connected layer to reduce the dimension, followed by a ReLU and a dropout layer. The first encoder layer of the Transformer model uses the embedded feature vectors as input, and the subsequent layers use the output tokens of the previous encoder layers. Each encoder layer is composed of a multi-head self-attention layer followed by a small feed-forward neural network.

Using the intermediate feature maps obtained from ResNet-101 [47] as input, a **Region Proposal Network (RPN)** generates bounding boxes for the objects proposed by the network. Multiple neural network layers are added to predict the corresponding class for each region and correct the bounding box for each of the proposed regions. Also, to implement geometric attention, the value of attention weight matrices changes: bounding box properties (such as center, width, and height) are combined with their corresponding attention weights using a high-dimensional embedding [116].

Liu et al. [77] introduce the **Global-and-Local Information Exploring-and-Distilling (GLIED)** approach that explores and distills the cross-modal source information. The Transformer-based structure globally captures the inherent spatial and relational groupings of the individual image regions and attribute words for an aspect-based image representation. Afterward, it extracts fine-grained source information locally for precise and accurate word selection. They used the RCNN-based visual features provided by Anderson et al. [7] for image regions extracted by Faster R-CNN [102].

Cornia et al. [23] introduce a fully attentive model called M^2—a Meshed Transformer with Memory for Image Captioning. The architecture is inspired by the Transformer model for machine translation and learns a multi-level representation of the relationships between image regions integrating learned *a priori* knowledge. The model incorporates two novelties: (1) image regions and their relationships are encoded in a multi-level fashion, in which both low-level and high-level relations are considered. The model learns and encodes *a priori* knowledge using persistent memory vectors. (2) The sentence generation—done with a multi-layer architecture—exploits both low- and high-level visual relationships via a learned gating mechanism, which weights multi-level contributions at each stage. This creates a mesh-like connection between the encoder and decoder layers. The encoder is in charge of processing regions in the input image and the relationships between them. Simultaneously, the decoder reads the output of each encoding layer and generates the caption word-by-word. All interactions between word- and image-level features are modeled via scaled dot-product attention without using recurrence.

Huang et al. [59] used a Transformer-like encoder paired with an LSTM decoder. Li et al. [67] investigated a Transformer-based sequence modeling framework named "ETA-Transformer." They have proposed **EnTangled Attention (ETA)**, which enables the Transformer to benefit from both semantic and visual information simultaneously. Liu et al. [80] introduce **CaPtion TransformeR (CPTR)**, which takes sequentialized raw images as input to the Transformer. As an encoder-decoder framework, CPRT is a full Transformer network that replaces the commonly used CNN in the encoder part with the Transformer encoder. A purely Transformer-based architecture, PureT, is designed by Wang et al. [123]. In PureT, SwinTransformer [82] replaces Faster-RCNN, and the architecture features a refining encoder and decoder.

Fang et al. [30] introduce a fully **VIsion Transformer-based image CAPtioning mode (ViTCAP)** along with a lightweight **Concept Token Network (CTN)**, which is used to produce concept tokens. The structure uses a vision transformer backbone as the stem image encoder, which produces grid features. CTN is then applied to predict semantic concepts. A multi-modal module uses grid representations and Top-K concept tokens as input to perform the decoding process.

Pseudo ground-truth concepts are extracted from the image captions using a simple classification task, and CTN is optimized to predict them during training.

Li et al. [74] designed a Transformer-style encoder-decoder structure called **Comprehending and Ordering Semantics Networks (COS-Net)**. A CLIP model (image encoder and text encoder) [99] is used as a cross-modal retrieval model that retrieves sentences semantically similar to the input image. The semantic words in retrieved sentences are treated as the primary semantic cues. A novel semantic *comprehender* is also introduced by the authors, which removes the irrelevant semantic words in those primary cues and simultaneously infers missing words visually grounded in the image. Afterward, a semantic ranker sorts the semantic words in linguistic order. Zeng et al. [141] propose a **Spatial-aware Pseudo-supervised (SP)** module that uses a number of learnable semantic clusters to quantize grid features with multiple centroids without direct supervision. These centroids aim to integrate grid features of similar semantic information together. In addition to the SP module, a simple weighted residual connection is introduced, named **Scale-wise Reinforcement (SR)** module. This module explores both low- and high-level encoded features concurrently.

Nguyen et al. [89] present a Transformer-only neural architecture titled **GRIT (Grid- and Region-based Image captioning Transformer)** that uses DETR-based detector along with grid- and region-based features. Hu et al. [54] have proposed ExpansionNet v2, which utilizes a novel technique titled Block Static Expansion layer. This technique processes the input by distributing it over a collection of sequences with different lengths, which helps to explore the possibility of performance bottlenecks in the input length in Deep Learning methods. This layer is designed to improve the quality of features refinement and ultimately increase the effectiveness of the static expansion. The architecture of ExpansionNet v2 follows the standard encoder-decoder structure and is implemented on top of Swin-Transformer [82].

2.6 Combining Transformers and Scene Graphs

A number of the works have experimented with model designs that incorporate both Transformers and scene graphs.

He et al. [48] aimed to employ the spatial relations between detected regions inside an image. In their proposed model, each Transformer layer implements multiple sub-transformers to encode relations between regions and decode information. The encoding method combines a visual semantic graph and a spatial graph. In another architecture introduced by Chen et al. [17], the encoder consists of two sub-encoders for visual and semantic information. Faster-RCNN proposes image regions, and a scene graph is built using the detected regions. GCN is then used to enrich the graph representation. A semantic matrix is learned from the scene graph and fed into a multi-modal attention module in the decoder. This module is used to leverage multi-modal representation in caption generation. Yang et al. [133] have proposed an architecture called ReFormer, which generates features with relation information embedded. ReFormer explicitly expresses the pair-wise relationships between objects present inside an image. ReFormer combines scene graph generation and image captioning using one modified Transformer model.

2.7 Vision Language Pre-training Methods

Some recent works have attempted pre-training paradigms to lessen the reliance of the models on fully supervised learning. A large-scale model is pre-trained on a dataset with an enormous amount of data by self-supervised learning. The pre-trained model is then generalized to various downstream tasks.

One widely used pre-trained model is **CLIP (Contrastive Language-Image Pre-Training)** [99]. CLIP is designed to provide a shared representation for both image and text prompts [88].

It has been trained on numerous images and captions using a contrastive loss, allowing for more consistency and correlation between its visual and textual representations. One of the recent works utilizing CLIP in the proposed method is ClipCap by Mokady et al. [88]. The authors introduce a model that produces a prefix for each caption by applying a mapping network over the CLIP embeddings. Next, a pre-trained language model (GPT-2 [100]) is fine-tuned to generate captions. This approach is inspired by Li et al. [72], who discussed the possibility of adapting a language model for new tasks by concatenating a learned prefix. Barraco et al. [12] investigate the role of CLIP [99] features in image captioning by devising an architecture composed of an encoder-decoder Transformer architecture.

Hu et al. [56] present the **VIsual VOcabulary pre-training (VIVO)**, which aims to learn a joint presentation of visual and text input. Unlike existing VLP models, which use image-caption pairs to pre-train, VIVO uses image-tag pairs for pre-training. In the pre-training stage, an image captioning model first learns to label image regions using image-tag pairs as training data. In the fine-tuning stage, the model learns to map an image to a sentence conditioned on the detected objects using image-caption pairs and their corresponding object tags. The sentences are learned from image-caption pairs, while object tags may refer to novel objects that do not exist in image-caption pairs. The addition of object tags allows for zero-shot generalization to novel visual objects for image captioning. Xia et al. [126] highlight that while recent pre-training methods for **vision-language (VL)** understanding tasks have achieved state-of-the-art performance, they cannot be directly applied to generation tasks. Xia et al. present **Cross-modal Generative Pre-Training for Image Captioning (XGPT)**, which uses a cross-modal encoder-decoder architecture and is directly optimized for generation tasks.

Li et al. [71] have proposed a pre-training method that leverages salient objects, which are usually present in both image and caption as anchor points. The method uses object tags as anchor points to align image and language modalities in a shared semantic space. The training samples are defined as triplets, each consisting of a word sequence, a set of object tags, and a set of image region features. This pre-training method can be applied to many vision-language tasks, including image-text retrieval, **Visual Question Answering (VQA)**, and image captioning. Many vision-language pre-training methods, including Reference [71], are built upon **Bidirectional Encoder Representations from Transformers (BERT)** [26]. These models use a two-stage training scheme in which the model first learns the contextualized vision-language representations by predicting the masked words or image regions based on their intra-modality or cross-modality relationships on large amounts of image-text pairs.

To counteract the problem of pre-training a single, unified model that is applicable to a wide range of vision-language tasks via fine-tuning, Zhou et al. [146] have introduced a new pre-training method for a unified representation for both encoding and decoding. The unified encoder-decoder model, called the **Vision-Language Pre-training (VLP)** model, can be fine-tuned for both vision-language generation (e.g., image captioning) and understanding tasks (e.g., visual question answering). This model uses a shared multi-layer Transformer network for encoding and decoding, which is pre-trained on large amounts of image-caption pairs. The VLP model is optimized for two unsupervised vision-language prediction tasks: bidirectional and **sequence-to-sequence (seq2seq)** masked language prediction. These two tasks only differ in what context the prediction conditions are on, which is controlled by specific self-attention masks for the shared Transformer network. The context of the masked caption word, which is the target of prediction, consists of all the image regions and all words on its right and left in the caption in bidirectional prediction. In contrast, in the seq2seq task, the context consists of all the image regions and the words on the left of the to-be-predicted word in the caption.

Li et al. [66] present mPLUG, a novel vision-language foundation model designed for both cross-modal understanding and generation. mPLUG aims to counteract some of the problems commonly witnessed in pre-training models, such as low computational efficiency and information asymmetry by novel cross-modal skip-connections. These skip-connections generate inter-layer shortcuts that skip a specific number of layers. This method is used to improve the slow full self-attention on the vision side. Liu et al. [78] present Prismer, a vision-language model that uses a group of domain experts to combine their knowledge and apply it to different vision-language reasoning tasks. Prismer performs well in fine-tuned and few-shot learning, while requiring significantly less training data compared to other models.

2.8 Unsupervised Methods and Reinforcement Learning

There has been a recent trend toward relaxing the reliance on paired image-caption datasets for image captioning. Many of the current research works employ reinforcement learning methods due to their unsupervised nature. The interest in unsupervised methods stems from the problem of the models relying almost entirely on the quality and volume of the image-caption pairs in datasets.

One early work by Gu et al. [42] involved generating captions in a pivot language and translating the caption to a target language. This method requires a paired image-caption dataset for the pivot language but does not use a paired dataset with captions being in the target language. Another research paper in this field used reinforcement learning with gradient policy along with RNNs in 2016 [101]. Shetty et al. [107] proposed the first study that explored using conditional GANs [87] to generate human-like and diverse descriptions.

Feng et al. [31] use a set of images, a sentence corpus, and a visual concept detector for unsupervised training. The images and the sentence corpus are projected into a common latent space such that they can reconstruct each other. The sentence corpus is prepared using the captions available on Shutterstock [109], which is a photo-sharing platform. On this platform, each image is uploaded with a caption. This corpus is not related to the images and is independent. The proposed structure comprises an image encoder, a sentence generator, and a discriminator. The Inception-V4 [112] is used as the image encoder, and the sentence generator and discriminator are both LSTMs [51].

Since no image-caption pairs exist, three new metrics have been introduced as three discriminators to evaluate the model's performance. The discriminator first distinguishes a real sentence from the sentence corpus from a sentence generated by the model, and the generator is rewarded at each timestep. By maximizing this reward, the generator tries to produce plausible sentences. However, more than this discriminator is needed, since the quality of the generated sentence and its relevance to the image must also be evaluated. To do so, the model must learn the visual contents of the image. The generated words are rewarded if the generated caption contains words whose corresponding visual concept is detected inside the image. This reward is called a "concept reward." Finally, since the performance of the model is much dependent on the performance of the visual concept detector and these detectors only detect a limited number of objects, images and captions are projected into a common latent space such that they can reconstruct each other.

Chen et al. [15] proposed an image captioning framework based on conditional generative adversarial nets as an extension of the reinforcement learning-based encoder-decoder architecture. Highlighting that the conventional encoder-decoder structures directly optimize one metric, which cannot guarantee improvement in all metrics, the paper designed a discriminator network to decide if a caption is human-described or machine-generated based on the idea of GANs. Two discriminator models have been designed and tested: a CNN-based discriminator model that uses the conditional CNN for real or fake sentence classification, and an RNN-based discriminator model that consists of the standard LSTM [51], a fully connected linear layer, and a softmax output

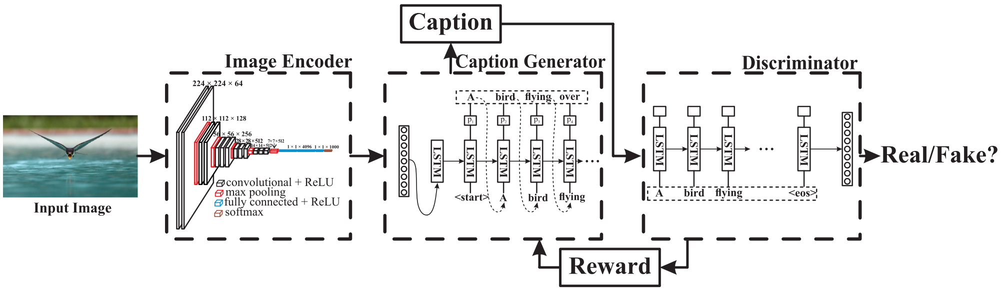


Fig. 4. The general workflow of unsupervised methods.

layer. The CNN-based framework was shown to improve the performance more than the RNN-based framework, while the RNN-based framework can save 30% training time. It was finally concluded that the ensemble results of four CNN-based (denoted as CNN-GAN) and four RNN-based (denoted as RNN-GAN) models could noticeably improve the performance of a single model.

Liu et al. [81] have introduced an image captioning module and a self-retrieval module. A Convolutional Neural Network extracts image features, and an LSTM [51] decodes a sequence of words based on these features. The self-retrieval module evaluates the similarity between the generated captions, the input image, and some "distractors." If the caption generator module generates distinct and proper captions, then the relevance between these captions and their corresponding images must be more than the relevance between the generated captions and unrelated, distracting images. This condition is represented as the text-to-image retrieval error and improves the performance of the image captioning module with back-propagation and the REINFORCE algorithm.

Guo et al. [45] used a discriminator structure similar to that of Reference [31]. The discriminator distinguishes whether the generated sentence is real and rewards the learner based on how real the sentences seem. Another discriminator distinguishes the style of the generated captions. Also, the LSTM [51] decoder used in Reference [41] has been used as a reinforcement learning agent making an action (prediction of the next word). After a sentence is completed, the agent will observe a sentence-level reward and update its internal state.

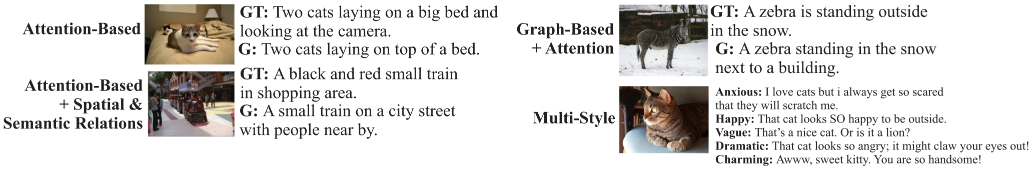
A block diagram of the general workflow of the unsupervised methods is shown in Figure 4. VGGNet [111] has been used as the image encoder, and the caption generator is an LSTM network [51]. Therefore, the overall design follows the typical encoder-decoder structure. The discriminator is also an LSTM network, which determines if the given caption is real (from the sentence corpus) or generated by the model. The generator is rewarded accordingly by the discriminator.

Taking the issues related to supervised settings into account, such as the tedious process of dataset preparation and the difficult training process, the unsupervised setting has been the focus of many recent works and is expected to become more favored in the future as well.

2.9 Generating Multi-style Captions

The papers discussed so far generate captions with a neutral tone. These generated captions usually describe factual data about image contents. Meanwhile, humans use many styles and tones in their daily speech to communicate with one another. Some of these styles and tones are humorous, hostile, and poetic. Incorporating these styles can help humans interact with the caption more and make the captions more attractive. Stylized captions can also be used in applications such as photo-sharing and Chatbots.

Shuster et al. [108] have added tone and style as a feature to their dataset, as well as images and their appropriate captions. This paper has introduced a novel structure called TransResNet,



GT: Two cats laying on a big bed and looking at the camera.
G: Two cats laying on top of a bed.

GT: A black and red small train in shopping area.
G: A small train on a city street with people near by.

GT: A zebra is standing outside in the snow.
G: A zebra standing in the snow next to a building.

Anxious: I love cats but i always get so scared that they will scratch me.
Happy: That cat looks SO happy to be outside.
Vague: That's a nice cat. Or is it a lion?
Dramatic: That cat looks so angry; it might claw your eyes out!
Charming: Awww, sweet kitty. You are so handsome!

Fig. 5. Sample captions generated by multiple methods in different categories. "GT" indicates "Ground Truth Caption," and "G" indicates "Generated Caption." The captions are generated by Huang et al. [59] (top-left), Wang et al. [121] (top-right), Li et al. [71] (bottom-left), and Shuster et al. [108] (bottom-right).

which projects images, captions, and their corresponding personality traits into a shared space using an encoder-decoder framework. Two classes of models have been considered: retrieval models and generative models. The retrieval model considers any caption in the entire dataset as a possible candidate response, whereas the generative model produces captions word-by-word via the aforementioned structure. The retrieval model has given better results.

A structure consisting of five modules for caption generation in different styles has been introduced [45] by Guo et al. The first module is a plain image encoder. Next is a caption generator module that outputs a sentence conditioned on a specific style. The following module is a caption discriminator that distinguishes a real sentence from a generated sentence. This discriminator is trained in an adversarial manner to encourage the learner to generate more convincing captions closer to the human language. Afterward, a style discriminator module that determines the style of the generated caption is used. Inspired by the fact that there is some content consistency between neutral captions and stylized captions, another module called "The Back-Translation Module" is also used. This module translates a stylized caption into a neutral one. (If a stylized caption is generated and translated to a factual and neutral caption, then we should arrive at the real factual caption.) This process is implemented using multi-lingual **neural machine translation (NMT)**, in which the stylized captions are considered input and neutral captions are considered output.

A figure consisting of some example captions from Sections 2.1, 2.1.1, 2.3, and 2.9 in this survey is shown in Figure 5. Each row belongs to a specific category in which two images are displayed, along with the captions describing them. For each image, the ground-truth caption and the generated caption are shown.

3 PROBLEMS IN IMAGE CAPTIONING

In image captioning, researchers are usually confronted with a set of problems, some of which commonly experienced in many artificial intelligence tasks such as the exposure bias problem [101], the loss-evaluation mismatch problem [41, 79, 103, 130], the vanishing gradient problem [50], and the exploding gradient problem [39, 95]. In addition, image captioning poses certain challenges unique to the task. These challenges include object hallucination, illumination conditions, contextual understanding, and referring expressions. We review some of the continuing problems in image captioning that come, in fact, as a part of the nature of the task.

3.1 Object Hallucination

Object hallucination [104] is a persistent problem for image captioning models, wherein the model detects objects that are not present in the input image. This can lead to poor performance in visually impaired users, who require accurate and concise captions. According to a study by MacLeod et al. [84], for many visually impaired people who prefer correctness over coverage, hallucination is a severe disadvantage for a captioning model and an obvious concern. Furthermore, object hallucination indicates an internal issue of the model. Rohrbach et al. have proposed a new metric to measure object hallucination, **CHAIR (Caption Hallucination Assessment with Image Relevance)**, which measures the proportion of generated words that correspond to objects in the input image according to ground truth sentences and object segmentations. The CHAIR metric has both per-instance and per-sentence variants, denoted as $CHAIR_i$ (Equation (9)) and $CHAIR_s$ (Equation (10)), respectively.

$$CHAIR_i = \frac{|\text{Hallucinated objects}|}{|\text{All objects mentioned}|} \tag{9}$$

$$CHAIR_s = \frac{|\text{Sentences with hallucinated objects}|}{|\text{All sentences}|} \tag{10}$$

According to the study performed by Rohrbach et al., models that perform better on standard evaluation metrics (such as BLEU [94] and SPICE [6]) perform better on CHAIR. However, this is not always true. It was found that the models that were optimized for CIDEr frequently hallucinated more. Also, models with attention tended to perform better on the CHAIR metric than models that did not incorporate attention. However, this gain was primarily due to these models' access to the underlying convolutional features and not the actual attention mechanism. Also, GAN-based models decreased hallucination, implying that GAN loss is beneficial in decreasing hallucination. This is due to the fact that the GAN loss encourages sentences to resemble human-generated captions. The presence of a hallucinated object likely suggests that a sentence is generated, and the discriminator dismisses the caption containing the hallucinated object.

3.2 Illumination Conditions

Illumination conditions are a critical factor that can impact the accuracy and reliability of the generated captions, particularly when the image is captured in low-light conditions or indoors. Poor lighting can result in images with reduced contrast, making it difficult for the captioning model to discern fine details and recognize objects, people, or scenes. Moreover, the presence of shadows or uneven illumination can further hinder the model's ability to accurately analyze the visual content. Shadows and uneven illumination can also further complicate the model's analysis of visual content. For example, an image of a black cat in a dimly lit room with uneven illumination may be difficult for the captioning model to recognize as a cat. To overcome these challenges, researchers have been actively exploring various techniques to improve the visual quality of the images [3, 4], including contrast enhancement, color correction, and low-light image enhancement. These techniques aim to mitigate the challenges posed by poor illumination and improve the accuracy of generated captions.

3.3 Contextual Understanding

Image captioning models also require the ability to understand the context of the scene, including the relationships between objects, the spatial arrangement, and the overall atmosphere [34, 125, 137, 138, 147]. This contextual understanding can be difficult to achieve, as it requires the caption generation model to have a deep understanding of the visual content and the ability to perform reasoning given the visual content.

3.4 Referring Expressions

Another problem in image captioning is the use of referring expressions, such as "the girl with the red hair" or "the dog in the corner." These expressions require the captioning models to identify and link the appropriate objects in the image. This can pose a challenge, especially if the objects are partially obscured or if there are multiple similar objects in the scene, and requires a combination of visual and linguistic understanding [21]. Referring expressions are important for improving caption quality, as they provide more detailed and informative descriptions of the objects in the image, allowing the model to generate more accurate and nuanced captions.

4 DISCUSSION

This section provides a comprehensive critical analysis of the methods falling in the different categories overviewed in Section 2. Each method—inevitably—possesses advantages and disadvantages. Nevertheless, considering these characteristics aids researchers in adopting a suitable solution. The technical details of the structures and methods discussed in this section have been explained in Section 2.

4.1 Using Attention

Attention-based methods attempt to imitate the human attention mechanism by showing the model "where to look at" during the training process. Attention is widely used in encoder-decoder architectures, where CNNs are typically used in combination with LSTMs to produce a representation for the given image and generate captions, respectively. Some of the papers focusing on attention-based methods have mentioned low precision in region selection for attention as a flaw of the attention-based methods. They claim that most of the attention-based methods presented choose regions of the same size and shape without considering image contents. They have also mentioned that determining the optimal number of region proposals will bring about an unresolvable tradeoff between small or large amounts of detail (or representing the image coarsely or finely).

One solution to this problem was proposed by Anderson et al. in Reference [7] as the "bottom-up and top-down" method. Another problem of the attention-based methods is the "single-stage" structure. Most of these methods are only a single encoder-decoder attention structure, which cannot provide rich captions for the images. In the multi-stage coarse-to-fine structure proposed by Gu et al. [41], at each stage, attention weights and hidden vectors generated by the previous stage decoder are used as input to the next stage decoder, reducing ambiguity in the captions. This structure allows for a richer caption at each stage.

Another problem associated with attention-based methods for image captioning is that a proper correlation between the vectors obtained from attention and caption is not guaranteed, and it might lead to improper results. If feature vectors do not contain valuable information, then the attention model still generates a vector that is a weighted sum over candidate vectors and is unrelated to the correct caption. To solve this issue, Huang et al. [59] have introduced an attention-based structure (Attention on Attention—or AoA) that contains one more level of attention. The authors have compared AoA with LSTM [51] and GRU networks [19]: Internal states, memories, and gates are used in LSTMs and GRUs to implement the attention mechanism. AoA only performs two linear transformations and does not require hidden states, making it computationally reasonable while outperforming LSTM. The combination of LSTM and AoA has been reported to be unstable, since it can reach a sub-optimal point. This means that increasing the volume of the stack and the number of gates to improve the performance is futile. Jiang et al. [61] state that the existing encoder-decoder models employ only one kind of CNN to describe image content. Consequently, the image contents will be described from only one specific viewpoint, and the semantic meaning of the input image cannot be comprehensively understood, which will restrict the performance. To improve the image captioning model, the model introduced by Jiang et al. [61] extracts diverse representations from multiple encoders. The novel **recurrent fusion network (RFNet)** proposed in the paper uses multiple CNNs as encoders. Each representation extracted from an individual CNN can act as an individual view of the image content.

4.1.1 Injecting Spatial and Semantic Relation Information into Attention-based Methods.

One of the significant downsides of the methods that only use the attention mechanism as their main solution for image captioning is that these methods fail to consider the spatial and semantic relations between image elements. Spatial and semantic relations in an image are integral to comprehension of the image contents [34, 125, 137, 138, 147]. For example, spatial relations in an image could help differentiate between "a person riding a horse" and "a person standing on a horse's back." Also, relative size can help differentiate between objects with their most significant difference being their size, such as violins and cellos. In addition to that, incorporating these relations makes the object detection task more precise.

As a possible solution, Herdade et al. [49] have introduced the "Object Relation Transformer." Pan et al. [93] have mentioned another problem about the attention-based image captioning methods: In most of these methods, only the first-order interactions between objects inside the image are observed. Their paper has claimed that, since the image captioning problem involves multi-modal data (image and text), multi-modal reasoning is needed, and observing the first-order interaction between features only will render more in-depth reasoning impossible. The structure proposed by Pan et al. uses spatial and channel-wise bilinear attention to extract second-order interactions.

Liu et al. [77] state that there is still great difficulty in deep image understanding, because the systems tend to view one image as unrelated individual segments and are not guided to comprehend the relationships between the objects inside the image. They argue that such understanding requires adequate attention to correlated image regions and coherent attributes of interest. To do so, they have presented the **Global-and-Local Information Exploring-and-Distilling (GLIED)** approach. Cornia et al. [22] claim that an attention-based architecture implicitly selects which regions to focus on, but it does not provide a way of controlling which regions are described and what importance is given to each region. The model suggested in their paper is able to focus on different regions in different orders following a given condition. Words can be organized into a tree-like structure, and a higher level of abstraction can be recovered considering the syntactic dependencies between words.

4.2 Using Graphs for Spatial and Semantic Relations

Graphs have been used extensively in many image captioning methods due to their ability to cohesively represent the relation between multiple elements. These methods have utilized graphs in two ways: scene graphs extracted from images and scene graphs extracted from textual data. Scene graphs have been used as a component inside encoder-decoder-based or unsupervised frameworks, and some have employed scene graphs along with Transformers. Graph-based methods pose challenges of their own. Yang et al. [134] rightfully state that an ever-present problem has never been substantially resolved: The different variants of the encoder-decoder-based framework, when fed an unseen image, usually produce a simple and trivial caption about the salient objects in the image, which is no better than a list of object detection. The model presented by Yang et al. adds the inductive bias of language generation to the encoder-decoder framework and uses scene graphs to connect the image and text modalities. Gu et al. [43] argue that the majority of image captioning studies are conducted in English, and preparing image-caption paired datasets in other languages requires human expertise and is time-consuming. The method introduced in their paper uses scene graphs as an intermediate representation of the image and sentence and maps the scene graphs in their feature space using cycle-consistent adversarial training.

4.3 Using Attention and Graphs

Considering how mutual correlations or interactions between objects are the natural basis for image description, Yao et al. [136] study the visual relationships between objects and how they can be utilized for this matter. They have built semantic and spatial correlations on image regions and used Graph Convolutions to learn richer representations. One major challenge of image captioning is the problem of grounded captioning. Most models do not focus on the same image regions as a human would while observing an image, which may lead to object hallucination [104].

Zhong et al. [145] addressed this problem by revisiting the representation of image scene graphs. The key idea is to select essential sub-graphs and only decode a single target sentence from a chosen sub-graph. The model can link the decoded tokens back into the image regions, demonstrating noticeable results for caption grounding. Another downside of the attention-based methods is that they do not incorporate the regions previously attended by the attention model. These regions can be used in the module's following region selections. Wang et al. [121] have integrated this point as well as the semantic relations between image elements in their proposed structure, which uses a novel content-based attention framework to store previously attended image regions. Chen et al. have discussed [18] that even though some methods focus on controlling expressive styles or attempt to control the description contents (discussed in Section 2.9), they can only handle a coarse-level signal. Their method uses a directed graph consisting of three node types grounded in the image, which allows for incorporating user intentions.

Li et al. [70] argue that most methods that devise semantic concepts treat entities in images individually and lack helpful, structured information. Therefore, they have utilized scene graphs along with CNN features from the bounding box offsets of object entities. Another work by Xu et al. [129] also addresses the lack of structured information in current systems. The authors propose the **Scene Graph Captioner (SGC)**, which is divided into three major components: the graph embedding model, the attention extraction model, and the language model. The attention extraction model is inspired by the concept of *small world* in the human brain. The work proposed by Lee et al. [65] uses visual relations from scene graphs directly instead of GCNs, claiming that it will avoid expensive graph convolutions. While the performance of some GCN-based models is slightly better, evading graph convolutions may be reasonable in some frameworks.

4.4 Using Convolutional Network-based Methods

LSTM networks [51] have been considered the standard for vision-language tasks such as image captioning and visual question answering due to their impressive ability to memorize long-term dependencies through a memory cell. However, training such networks can be considerably challenging due to the complex addressing and overwriting mechanism combined with the required processing being inherently sequential, and the significant storage required in the process. LSTMs [51] also require more careful engineering when considering a novel task [8]. Earlier, CNNs could not perform as well as LSTMs on vision-language tasks. The recent advances in convolutional structures on other sequence-to-sequence tasks have enabled researchers to use CNNs in many other vision-language tasks. Also, CNNs produce more entropy [8], which can be helpful for diverse predictions, have better classification accuracy, and do not suffer from vanishing gradients. Aneja et al. [8] proposed a convolutional model that uses masked convolutions instead of LSTM or GRU units. This work also experimented with attention by forming an attended image vector and adding it to the word embedding at every layer. Doing so, the model has outperformed the attention baseline [128]. With attention, the model could identify salient objects for the given image. Arguing that RNNs or LSTMs [51], which are widely used in image captioning, cannot be computed in parallel and also ignore the underlying hierarchical structure of the sentences, Wang et al. [122] designed a framework entirely relying on CNNs. The proposed model can be computed in parallel and is faster to train. However, convolutional network-based methods still need improvement in terms of performance.

4.5 Using Transformers

The encoder-decoder framework continues to dominate the image captioning world, with the models only varying in details and sub-modules. The recent success of Transformers in natural language processing tasks has inspired many researchers to replace the RNN model with Transformer in the decoders, aiming to benefit from its excellent performance and the possibility of parallel training. Transformers have been the center of attention in the computer vision field as well, with models such as DETR [14], ViT [28], SETR [144], and IPT [16]. Liu et al. [80] have proposed the **CaPtion TransformeR (CPTR)**, a full Transformer network to replace the widely used CNN in the encoder part of the encoder-decoder framework. Fang et al. [30] criticize the use of object detectors as a tool to provide visual representation, stating that it may lead to heavy computational load and that they require box annotations. Fang et al. [30] have introduced the detector-free ViTCAP model with a fully Transformer architecture, which uses grid representations without regional operations.

Nguyen et al. [89] mention another issue with CNN-based detectors. CNN-based detectors use **non-maximum suppression (NMS)** at the last stage of computation to remove redundant bounding boxes. As a consequence, end-to-end training of an entire model consisting of detector and decoder modules becomes difficult. To overcome this problem and to reduce their high computation cost, Nguyen et al. employ the Deformable DETR [149] and replace the CNN backbone in the original design with Swin Transformer. The **COS-Net model (Comprehending and Ordering Semantics Network)** proposed by Li et al. [74] aims to unify semantic comprehending and ordering. COS-Net uses a CLIP model (image encoder and text encoder) [99] as a cross-modal retrieval model that retrieves sentences semantically similar to the input image.

Zeng et al. [141] argue that directly operating at grid features may lead to the loss of spatial information caused by the flattening operation. The objective of the **Pseudo-supervised (SP)** module designed by the authors is to resolve this issue. Also, the **Scale-wise Reinforcement (SR)** module has been introduced to maintain the model size and improve performance. Wang et al. [123] argue that using a network such as Faster-RCNN as the encoder divides the captioning task into two stages and thus limits it. The PureT model built by the authors is a pure Transformer-based structure that integrates the captioning task into one stage and enables end-to-end training.

ExpansionNet v2 introduced by Hu et al. [54] aims to solve the problem of performance bottlenecks in the input length in Deep Learning methods for image captioning. To address this issue, the authors introduce a new technique called Block Static Expansion, which distributes and processes the input over a collection of sequences with different lengths. This method helps to improve the quality of features refinement and ultimately increase the effectiveness of the static expansion.

4.5.1 Using Graphs and Transformers. Some current captioning encoders use a GCN to represent the relation information. Yang et al. [133] highlight that these encoders are ineffective in image captioning due to the use of methods such as Maximum Likelihood Estimation rather than a relation-centric loss and the use of pre-trained models to obtain relationships instead of the encoder to improve model explainability. Yang et al. propose the ReFormer architecture, which applies the objective of scene graph generation and image captioning by means of one modified Transformer. Chen et al. [17] use the Transformer as their base architecture in the model **SGGC (Scene Graph Guiding Captioning)**. The encoder is composed of two sub-encoders named visual encoder and semantic encoder. In the visual encoder sub-component, a Transformer encoder consisting of N identical encoding layers has been used instead of the general CNN-based encoder to capture the relationships between visual regions better. Scene graphs have been used as additional guidance for decoder generation. While Transformers are suitable for self-supervised pretext tasks on large-scale data, training can become expensive and burdensome. There is a need for more economic Transformer-based large-scale multi-modal models that can be achieved by means of incorporating more inductive bias about vision and language data [131].

4.6 Using Vision-language Pre-training for Image Captioning

VLP has remarkably contributed to the recent advances in image captioning and is currently the dominant training method for VL tasks. In VLP approaches, a large-scale model is usually pre-trained on massive amounts of data using self-supervised learning and then generalized to adapt to downstream tasks. References [71, 143, 146] have extensively observed the effect of pre-training objective methods and architectures. The scale of the pre-training dataset is also believed to be a crucial factor in outstanding performance. VLP helps alleviate some of the problems experienced in conventional image captioning methods. The conventional methods typically need to minimize the gap between the visual and textual modals and are therefore resource-hungry [88]. Excessive training time and numerous trainable parameters are also required, reducing their practicality. However, given new samples, the models need to be updated to adapt to new inputs. This brings about the need for lightweight models with faster training times and fewer parameters.

It has recently been observed that powerful vision-language pre-trained models improve zero-shot performance dramatically and reduce training time. One such pre-trained model is CLIP [99]. Mokady et al. [88] have leveraged CLIP encoding as prefix to the captions in their ClipCap model. A lightweight Transformer-based mapping network is trained from the CLIP embedding space and a learned constant. The GPT-2 network is used as the language model to generate captions given the prefix embeddings. Taking note of the fact that salient objects are usually present in both image and the corresponding caption, the pre-training method proposed by Li et al. [71] leverages these objects as anchor points to tackle issues such as ambiguity and a lack of grounding. Zhou et al. presented a pre-training method for a unified representation for encoding and decoding in Reference [146]. The VLP model proposed in this paper has the advantage of unifying the encoder and decoder and learning a more universal contextualized vision-language representation, which can be fine-tuned for generation and understanding tasks easily. This unified procedure results in a single model architecture for the two distinct vision-language prediction tasks (bidirectional and seq2seq). This alleviates the need to train multiple pre-training models for different tasks without significant performance loss. To fine-tune for the image captioning task, the VLP model is fine-tuned on the target dataset using the seq2seq objective. Hu et al. [56] point out that while many VLP methods have been introduced that learn vision-language representations through training large-scale Transformer models, most are designed for understanding tasks. The few solutions that can be applied to image captioning [71, 146] use paired image-caption data for pre-training, which cannot improve zero-shot performance. VIVO, proposed by Hu et al., learns vision-language alignment on image-tag pairs. Since caption annotations are not needed, many existing vision datasets originally prepared for tasks such as image tagging or object detection can be used. Xia et al. [126] emphasize that VL generation tasks necessitate the ability to learn generation capabilities as well as the ability to understand cross-modal representations. Also, Xia et al. explain that the pre-trained models developed for understanding tasks only provide the encoder, and separate decoders need to be trained to enable generation. In addition to this deficiency, none of the pre-training tasks are designed for the whole sentence generation. The XGPT takes advantage of a cross-modal encoder-decoder architecture and is directly optimized for VL generation tasks. Three generative pre-training tasks have been designed to countervail the lack of pre-training objectives for generation tasks, namely: **Image-conditioned Masked Language Modeling (IMLM)**, **Image-conditioned Denoising Autoencoding (IDA)**, and **Text-conditioned Image Feature Generation (TIFG)**. Li et al. [66] mention computational inefficiency and information asymmetry as some of the shortcomings of existing pre-trained models. Li et al. [66] have proposed the mPLUG model, which incorporates a novel cross-modal fusion mechanism with cross-modal skip-connections to alleviate these problems. Liu et al. [78] point out the data insensitivity problem and heavy computations associated with current vision-language problem and take a different approach in their model Prismer to learn domain knowledge via distinct and separate sub-networks, referred to as experts. Prismer includes modality-specific experts that encode multiple types of visual information directly from their corresponding network outputs. The expert models are pre-trained and frozen individually and are connected through lightweight trainable components. This approach results in a significant reduction in total network parameters.

4.7 Using Unsupervised Methods and Reinforcement Learning

The research works discussed in the aforementioned categories used a combination of images and their corresponding captions to train the structures they introduced and generated captions for new images while optimizing metrics. Training these supervised methods is challenging and involves some problems. One problem is that most of the research on image captioning has only worked on generating captions in the English language, and a proper dataset consisting of captions in multiple languages is not available. Preparing such a dataset requires the skills of human experts and is very time-consuming. Preparing a dataset of images and their corresponding captions is generally a difficult task. The Microsoft COCO dataset [76], which is widely used in image captioning, is much smaller than other datasets specifically designed for the object detection task, such as ImageNet and Open Images [23]. Microsoft COCO dataset [76] has 100 object classes only; consequently, the models trained on this dataset fail to generalize for new images that were not covered in the dataset. A considerable part of image captioning research is moving towards unsupervised methods to solve these issues. The early works improved the diversity of the captions; however, they sacrificed overall performance.

Feng et al. [31] have used a sentence corpus, a visual concept detector, and a set of images for unsupervised training. The model is composed of an image encoder, a sentence generator, and a discriminator. The results obtained from this research work have been criticized by Gu et al. [43] (discussed in Section 2.2). It has been explained that considering the limitations imposed by supervised learning, this research work has not achieved significant results, and the performance of the proposed model is not satisfactory. Gu et al. [43] use an unsupervised method (CycleGAN) to align the scene graph and the captions. Chen et al. [15] point out one issue with conventional encoder-decoder structures: Many directly optimize one or a combination of metrics. This can not guarantee consistent improvement over all metrics. As a solution, Chen et al. have designed a discriminator network based on the idea of GANs, which judges if a caption is human-generated or produced by a machine. Liu et al. [81] have introduced a system consisting of a captioning module and a self-retrieval module. The notable part of this work is the self-retrieval module (which uses the REINFORCE algorithm) that improves the performance of the aforementioned structure while only training on partially labeled data.

4.8 Captioning in Multiple Styles

Some of the papers covered in this survey generate captions in multiple styles, with some of these styles being humorous or hostile. The structure called “TransResNet,” presented by Shuster et al. [108], considers two classes of models: retrieval and generative. While the retrieval model has given better results, a disadvantage of the retrieval models for caption generation is that these models do not produce a new caption and only choose a caption from a massive dataset. The retrieval models usually generate general and repetitive captions. This pushes many researchers to use unsupervised methods. Guo et al. [45] have stated that incorporating appropriate styles into captions will enrich their clarity and appeal and allows for user engagement and social interactions. The structure presented by Guo et al. is composed of five modules for caption generation in different styles.

Stylized captions can help improve user interaction. However, since neutral captions that report factual data are more appropriate for visually impaired individuals, stylized captions may not be the best choice to utilize in assistive technologies.

5 DATASETS AND PERFORMANCE COMPARISON

The methods discussed in previous sections use various datasets and are evaluated with multiple evaluation metrics. In this section, we review the datasets and metrics widely used in recent research works in depth. The available datasets for the image captioning task are still small compared to that of object detection, and the evaluation metrics have many limitations. Considering the increasing importance of the image captioning task, preparing richer datasets and more accurate metrics can be vital to the growth and improvement of the task.

5.1 Datasets Used by Recent Works

5.1.1 Microsoft COCO. The MS COCO dataset [76] is a vast dataset for object detection, image segmentation, and image captioning. This dataset contains many features, such as image segmentation, 328,000 images, 91 object classes, and 5 captions for each image.

5.1.2 Flickr30K, Flickr30K Entities, and FlickrStyle10k. The Flickr30K dataset [139] is introduced for the automatic image captioning and grounded language understanding task. This dataset contains 31,000 images collected from the Ficker website, along with 158k captions written by humans. This dataset contains a detector for everyday objects, a color classifier, and a bias toward selecting larger objects. The Flickr30K Entities dataset [97] is based on this dataset and contains 158k captions from Flickr30K with 244k coreference chains that link mentions of the same entities in images. The dataset also contains 276k manually annotated bounding boxes corresponding to each entity. The FlickrStyle10k dataset [33] contains 10k images with captions of varying styles. Training data consists of 7k images, and the testing and evaluation data consist of 2k and 1k images, respectively. Each image has captions in different styles, such as poetic, humorous, and neutral (factual).

5.1.3 Visual Genome. Unlike the other dataset discussed that only had one caption for the entire image, this dataset [63] presents a separate caption for each image region. This dataset comprises seven parts: region descriptions, object bounding boxes, attributes, relationships, region graphs, scene graphs, and question-answer pairs. The Visual Genome dataset contains more than 108k images, with each image having an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects.

5.1.4 TextCaps. This dataset [110] aims to help train visual assistants for visually impaired individuals, focusing on presenting captions for images with written text inside them. This dataset presents 145k captions for 28k images.

5.1.5 VizWiz-Captions. This dataset [46] has been introduced as a dataset appropriate for image captioning for visually impaired individuals. This dataset consists of 23,431 training images and 117,155 training captions, 7,750 validation images, 38,750 evaluation captions, 8,000 images, and 40,000 testing captions. The images have been taken directly by visually impaired individuals.

5.1.6 Google's Conceptual Captions. This dataset [106] consists of approximately 3.3 million images and captions. The images have been collected from the Internet first, along with the "alt-text" associated with them. These image-caption pairs have then been filtered and processed to

Table 1. Most Common Datasets

Dataset	Total Images	Objects/Image	Object Classes	Captions/Image
Visual Genome [63]	108,077	36.17	80,138	5.4m R.D.
MS COCO [76]	330,000	7.57	91	5
Flickr30K Entities [97]	31,783	8.7	44,518	5
OpenImagesV6:V.R. [40]	375,000	8.4	-	1
Flickr30K [139]	31,000	-	-	5
FlickrStyle10K [33]	10,000	-	-	2
OpenImagesV6:L.N. [98]	849,000	-	-	1
SentiCap [86]	3171	-	-	6
TextCaps [110]	28,408	-	-	5
VizWiz-Captions [46]	39,181	-	-	5
nocaps [1]	15,100	-	680	11
Conceptual Captions [106]	3 mil<	-	-	1

Details (R.D. Indicates "Region Descriptions," L.N. Indicates "Localized Narratives," and V.R. Indicates "Visual Relationships").

extract appropriate captions for the images that describe image contents. This dataset is split into training and evaluation splits. There are 3,318,333 image-caption pairs in the training split and 15,840 image-caption pairs in the evaluation split.

5.1.7 Nocaps. The **Novel Object Captioning at Scale (nocaps)** dataset [1] has been presented to encourage the development of captioning models that can surpass the limitation of visual concepts in existing datasets. The introduced benchmark is composed of 166,100 human-generated captions describing 15,100 images from the Open Images validation and test sets. The training data consists of Open Images image-level labels and object bounding boxes in addition to COCO image-caption pairs. Considering that Open Images contains many more object classes not present in COCO, about 400 object classes in test images have almost no associated training captions.

5.1.8 Open Images V6: Visual Relationships and Localized Narratives. The Open Images dataset [64] contains various sections for object detection, image segmentation, object relationships, and more. The dataset includes approximately 9 million images in 600 different classes. Each image contains an average of 8.4 objects. One section of this dataset is the Visual Relationships section, which contains 329 tertiary relationships for 375k images. The most recent version of this dataset is available at Reference [40]. In 2020 and the sixth version of the Open Images dataset, a new section under the name of "Localized Narratives" was added [98]. This new section contains 1 million and 671k images from the Open Images Dataset. A human describer has described each image in the dataset via a voice recording while moving their computer mouse on the regions they were describing. Since the words of the caption are in sync with the mouse movements, the location associated with each word is available.

The details regarding datasets discussed in this section have been summarized in Table 1.

5.2 Evaluation Metrics for Image Captioning Methods

The metrics discussed below fall into two categories: the *text evaluation* metrics and the *caption evaluation* metrics. The text evaluation metrics evaluate machine-generated text portions independently. Most of these metrics were introduced for evaluating the text generated by machine translation models. The caption evaluation metrics evaluate the captions generated by the models and have been designed specifically for the image captioning task.

5.2.1 BLEU (Bilingual Evaluation Understudy). BLEU [94] is an evaluation metric for machine-generated texts. Separate parts of a text are compared against a set of reference texts, and each part receives a score. The overall score is an average over these scores; however, the syntactical correctness is not evaluated. The performance of this metric varies based on the references used and the size of the generated text. The BLEU metric is a widely used metric due to it being a pioneer in evaluating machine-generated texts, being independent of language, their simplicity, high speed, low cost, and being quite comparable with human judgment. BLEU counts the consistent *n*-grams in the machine-generated text and the reference text. *n*-grams are a contiguous sequence of n items in a text in the field of computational linguistics and probability. These items can be phonemes, syllables, letters, words, or base pairs. The number "n" determines the number of grams that will be compared against each other. Usually, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are computed using the BLEU metric. To compute BLEU-n, the *n*-grams of 1 to "n" are computed, and each is assigned one single weight. Next, the geometric mean of these *n*-grams is calculated according to these weights. For example, when computing BLEU-4, the *n*-grams of 1 to 4 are calculated, and each is given the value 0.25 as their weight, followed by the geometric mean being computed over these values. This metric does have some disadvantages, such as the fact that the computed scores are only high when the generated text is short. Also, in some cases, a high score achieved using this metric is unreliable and does not mean a higher-quality text.

5.2.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE [75] is a set of metrics that evaluates the quality of text summarization. ROUGE determines the quality of a summary by comparing it to other ideal human-created summaries: the number of overlapping units, such as *n*-grams, word sequences, and word pairs between the machine-generated summary and the ideal summaries are counted. Multiple measures are introduced: ROUGE-N (which counts the overlap of *n*-grams between the machine-generated summary and the ideal summary; ROUGE-1 and ROUGE-2 are subsets of ROUGE-N), ROUGE-L (which is essentially longest common subsequence-based statistics and considers sentence-level structure similarity naturally and automatically finds the longest co-occurring in sequence *n*-grams), ROUGE-W (which is based on weighted longest common subsequence that prefers consecutive longest common subsequences), ROUGE-S (which is based on skip-bigram co-occurrence, with skip-bigram being any pair of words in their sentence order), and ROUGE-SU (which is based on skip-bigram plus unigram co-occurrence), with each being used in a specific application. This metric does not perform well for evaluating summaries in more than one text.

5.2.3 METEOR (Metric for Evaluation of Translation with Explicit Ordering). This metric [11] compares word segments against reference texts. This method is based on the harmonic mean of unigram precision and recall (recall is weighted higher than precision). METEOR has features such as stemming and synonymy matching in addition to the standard exact word matching. This metric makes a better correlation at the sentence level or segment level.

5.2.4 CIDEr (Consensus-based Image Description Evaluation). This metric [117] is explicitly designed for evaluating image captions and descriptions. In contrast to other metrics working with only five captions per image—which makes them unsuitable for evaluating the consensus between the generated captions and human judgments—CIDEr reaches this level of consensus using **term-frequency inverse document frequency (TF-IDF)**. CIDER is technically an annotation modality for automatically computing consensus. A measure of consensus encodes how often *n*-grams in the candidate sentence are present in the reference sentences. Also, *n*-grams not present in the reference sentences must not exist in the candidate sentences. Furthermore, lower weight must be given to *n*-grams frequently appearing across all images in the dataset,

Table 2. The Independent Results

Ref	B-1	Ref	B-2	Ref	B-3	Ref	B-4
Zhong et al. [145]	**90.7**	Li et al. [74]	**69.1**	Li et al. [74]	**54.9**	Zhong et al. [145]	**59.3**
Nguyen et al. [89]	84.2	Pan et al. [93]	66.8	Pan et al. [93]	52.6	Li et al. [66]	46.5
Hu et al. [54]	83.5	Liu et al. [80]	66.6	Liu et al. [80]	52.2	Li et al. [74]	42.9
Li et al. [74]	83.5	Jiang et al. [61]	64.7	Jiang et al. [61]	50.0	Hu et al. [54]	42.7
Yang et al. [133]	82.3	Li et al. [70]	63.2	Li et al. [70]	48.3	Nguyen et al. [89]	42.4
Cornia et al. [23]	82.0	Liu et al. [81]	63.1	Liu et al. [81]	48.0	Li et al. [71]	41.7
Liu et al. [80]	81.7	Gu et al. [41]	62.5	Gu et al. [41]	47.9	Pan et al. [93]	40.7
Pan et al. [93]	81.7	Chen et al. [17]	60.7	Wang et al. [121]	46.5	Cornia et al. [23]	40.5
Huang et al. [59]	81.6	Wang et al. [121]	60.3	Chen et al. [17]	46.2	Liu et al. [78]	40.4
Li et al. [67]	81.5	Aneja et al. [8]	55.3	Aneja et al. [8]	41.8	Huang et al. [59]	40.2

Top 10 Methods - BLEU-1, BLEU-2, BLEU-3, and BLEU-4 (B: BLEU [94], Ref: Reference).

Table 3. The Independent Results

Ref	M	Ref	R	Ref	C	Ref	S
Zhong et al. [145]	**40.1**	Zhong et al. [145]	**71.5**	Cornia et al. [22]	**209.7**	Cornia et al. [22]	**48.5**
Li et al. [66]	32.0	Hu et al. [54]	61.1	Chen et al. [18]	204.2	Chen et al. [18]	42.1
Liu et al. [78]	31.4	Li et al. [74]	61.0	Zhong et al. [145]	166.7	Zhong et al. [145]	30.1
Hu et al. [55]	31.4	Nguyen et al. [89]	60.7	Li et al. [66]	155.1	Li et al. [66]	26.0
Li et al. [74]	30.8	Fang et al. [30]	60.1	Hu et al. [55]	145.5	Hu et al. [55]	25.5
Nguyen et al. [89]	30.6	Barraco et al. [12]	59.9	Nguyen et al. [89]	144.2	Hu et al. [54]	24.7
Hu et al. [54]	30.6	Yang et al. [133]	59.8	Hu et al. [54]	143.7	Li et al. [74]	24.7
Li et al. [71]	30.6	Pan et al. [93]	59.7	Li et al. [74]	143.0	Li et al. [71]	24.5
Fang et al. [30]	30.1	Cornia et al. [23]	59.5	Li et al. [71]	140.0	Liu et al. [78]	24.4
Barraco et al. [12]	30.0	Liu et al. [80]	59.4	Barraco et al. [12]	139.4	Nguyen et al. [89]	24.3

Top 10 Methods (M: METEOR [11], R: ROUGE [75], C: CIDEr [117], S: SPICE [6], Ref: Reference).

since they are likely to contain less information. To encode this, Vedantam et al. [117] performed a TF-IDF weighting for each n-gram. A version of CIDEr called CIDEr-D exists as a part of the Microsoft COCO evaluation server.

5.2.5 SPICE (Semantic Propositional Image Caption Evaluation). The SPICE metric [6] is a metric for evaluating image captions based on semantic context. This metric measures how well objects, attributes, and the relations between them are covered in image captions. A scene graph is used to extract the names of different objects, attributes, and the relationships between them from image captions. The metric utilizes semantic representations produced by this graph.

The discussed methods are far from human judgment in terms of quality due to various factors. Using external knowledge databases along with evaluation metrics can help improve evaluation quality.

5.3 Comparing Independent Results

Many research works have reported their results independently, as well as the results reported by the Microsoft COCO servers.

In this section, we list the results reported independently by the works covered in this survey in Tables 2 and 3. We have listed the best results for the research works that reported results under different settings (for example, optimization using different loss functions). The best performances are highlighted with boldface font.

Among the research works covered in this survey paper, Reference [145] (which introduces a sub-graph proposal network along with an attention-based LSTM decoder) has had the best results in BLEU-1 (90.7) and BLEU-4 (59.3), as well as METEOR [11] (40.1) and ROUGE [75] (71.5), while Reference [74] (COS-Net, a model that uses CLIP image and text encoder as a cross-modal retrieval model) has had the best results in BLEU-2 (69.1) and BLEU-3 (54.9). Also, Reference [22] (Show, Control, and Tell) has achieved the best CIDEr and SPICE results (209.7 and 48.5, respectively).

A recurring pattern among the best-performing methods presented in Tables 2 and 3 is the application of Transformers, scene graphs, and vision language pre-training methods [12, 17, 18, 23, 30, 54, 55, 59, 66, 67, 70, 71, 74, 78, 80, 89, 93, 133, 145]. These methods owe their performance to the capabilities of Transformers, scene graphs, and vision language pre-training methods. Transformers are capable of capturing complex relationships between objects and their surroundings, making them particularly effective in handling long-range dependencies in image sequences. Scene graphs, however, represent the relationships between objects within an image and allow for efficient inference of the visual content. Another desirable feature of graphs is their ability to represent composite and unstructured data types, as they provide a flexible and efficient way to model the complex relationships and interconnections between various entities within a system. In addition to Transformers and scene graphs, some of the high-performing image captioning methods in Tables 2 and 3 utilize vision language pre-training techniques [12, 55, 66, 71, 74, 78]. These methods involve training a model on large datasets that consist of both visual and textual information, allowing the model to learn a joint embedding space. By pre-training on such datasets and acquiring knowledge from multiple modalities, the model can effectively learn to understand visual content and generate natural language descriptions. The integration of these techniques in captioning models has led to a notable improvement in their overall performance, as evidenced by the results presented in Tables 2 and 3.

6 CHALLENGES AND THE FUTURE DIRECTIONS

Despite the abundance of solutions and methods presented to solve the image captioning problem, some challenges and open problems remain. The performance of the supervised methods relies significantly on the quality of the datasets. However, datasets can not cover the real world regardless of how massive they are, and the applicability of supervised methods is limited to the set of objects the detector is trained to distinguish. However, datasets with image-caption pairs inevitably contain more examples of a specific situation (one example being: “man riding a skateboard”). These examples in the training data falsely bias the model towards generating more captions similar to those examples rather than including actual detected objects [70]. The supervised paradigm overly relies on the language priors, which can lead to the object-hallucination phenomenon as well [74].

The problems associated with the supervised methods encourage researchers to devise unsupervised techniques. However, due to the different properties of image and text modalities, the encoders of image and sentence cannot be shared. Therefore, the critical challenge in an unpaired setting is the gap of information misalignment in images and sentences [43]. The current unsupervised image captioning methods still need to catch up in performance rankings.

One promising direction of research is using scene graphs for image captioning. However, despite the many possibilities unveiled by scene graphs, discussed extensively in the previous sections, utilizing them comes with challenges. Constructing scene graphs is a complicated task in itself, and due to the interactions between objects being beyond simple pairwise relations, integrating scene graphs is quite tedious [129]. Also, scene graph parsers are still not as powerful

[120, 134]. According to some of the works that studied the impact of scene graphs on the quality of the captions, scene graphs are effective only if pre-training of the scene graph generators is done with visually relevant relation data [65].

VLP methods have been used to resolve some of the flaws with supervised methods and object detector-based designs. However, most VLP approaches are catered to understanding tasks, and generation tasks such as image captioning demand more capabilities. A number of the recent works covered in this article have aimed to fulfill this need. However, this field needs more investigation and analysis. Moreover, detector-free designs have a rising popularity. In these designs, the detector is removed for the vision-language pre-training in an end-to-end fashion [30]. Also, a general visual encoder replaces the detector and is used to produce grid features for later cross-modal fusion. However, the construction of a stronger detector-free image captioning model still needs investigation. Despite the challenges faced when working with scene graphs, vision language pre-training methods, and Transformers, almost each one of the best-performing models according to evaluation metrics uses one or a combination of these techniques, as shown in Section 5.3. This further proves the potential of these techniques in solving the image captioning problem and are promising tools for the future of generative tasks. Specifically, graphs are valuable in representing complex relationships and interconnections between different entities, particularly for composite, semi-structured, and unstructured data that may not be easily handled by other types of data models. Considering the recent advancements in generative artificial intelligence such as **large language models (LLMs)** [92] and **multimodal language models (MLLMs)** [5, 60, 68, 91], the need for representation methods capable of handling such data types will become more and more visible and felt in near future.

Another gap in the literature is the lack of focus on the application of image captioning for the visually impaired. Describing images can be the core of a vision assistant designed to aid the visually impaired in their daily lives: One can be informed of potential dangers in their environment and have a general understanding of what is happening around them. Considering the issues mentioned earlier and gaps, unsupervised learning and unpaired setting are of great potential. Also, the graph-based approach is expected to become even more popular in the near future. LLMs, MLLMs, and Transformers in combination with vision-language pre-training methods are also very likely to become standard practice.

7 CONCLUSION

This article has covered recent image captioning methods, provided a taxonomy of the approaches, and mentioned their features and properties. We also discussed the common problems in image captioning, reviewed datasets and evaluation metrics, compared the performance of the covered methods and algorithms in terms of experimental results, and highlighted the challenges and future directions in image captioning. Despite the numerous methods and solutions presented for the image captioning problem, there are still some major problems and challenges for which few solutions have been suggested. However, the generated captions still need to be higher in quality and are far from human-generated captions. Also, the datasets cannot cover the infinite real world. The evaluation metrics still need to be improved and are still not ideal for evaluating the precise performance of the models. However, VLP methods are frequently used in recent works and have shown promising performance. VLP methods and Transformers are likely to be inseparable components of models in the future of image captioning.

Moreover, more research needs to be done on visual assistants for visually impaired individuals. Preparing such an assistant requires certain features to be implemented, making it different from the other applications of image captioning. The best models presented by the research works do not perform well as visual assistants and do not consider the specific demands and needs of visually impaired people. A proper caption for a visually impaired person includes the most important aspects of the image first and the other noticeable details afterward. The surroundings and finer details must also be described, such as details about the textures and the position of objects relative to each other. Therefore, a caption appropriate for the needs of visually impaired individuals is denser and contains much more detail compared to the captions generated by conventional methods and models. Also, the caption generation process may be altered in a way that the initial caption provided to the user can be more general and shorter. The caption may become denser and more detailed upon the user asking more questions about the image. Considering the importance of the aforementioned issues and the growing number of visually impaired individuals, a noticeable lack of an efficient solution remains. Valuable research work in this field would be automatic image captioning with a particular focus on creating a visual assistant for visually impaired individuals.

REFERENCES

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19)*. 8948–8957.

[2] Hiba Ahsan, Daivat Bhatt, Kaivan Shah, and Nikita Bhalla. 2021. Multi-modal image captioning for the visually impaired. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 53–60.

[3] Rayan Al Sobbahi and Joe Tekli. 2022. Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: Overview, empirical evaluation, and challenges. *Sig. Process.: Image Commun.* 109, C (2022), 116848. https://dl.acm.org/doi/abs/10.1016/j.image.2022.116848

[4] Rayan Al Sobbahi and Joe Tekli. 2022. Low-light image enhancement using image-to-frequency filter learning. In *Proceedings of the 21st International Conference on Image Analysis and Processing (ICIAP'22)*. Springer, 693–705.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* 35 (2022), 23716–23736.

[6] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 382–398.

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 6077–6086.

[8] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. 2018. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5561–5570.

[9] Hareem Ayesha, Sajid Iqbal, Mehreen Tariq, Muhammad Abrar, Muhammad Sanaullah, Ishaq Abbas, Amjad Rehman, Muhammad Farooq Khan Niazi, and Shafiq Hussain. 2021. Automatic medical image interpretation: State of the art and future directions. *Pattern Recog.* 114, Article 107856 (2021).

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, Yoshua Bengio and Yann LeCun (Eds.).

[11] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[12] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4662–4670.

[13] John Adrian Bondy and Uppaluri Siva Ramachandra Murty. 1976. *Graph Theory with Applications*, Vol. 290. North-Holland.

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. 213–229.

[15] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. 2019. Improving image captioning with conditional generative adversarial nets. *Proc. AAAI Conf. Artif. Intell.* 33, 01 (2019), 8142–8150.

[16] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21).* 12294–12305.

[17] Haishun Chen, Ying Wang, Xin Yang, and Jie Li. 2021. Captioning transformer with scene graph guiding. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'21).* 2538–2542.

[18] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9962–9971.

[19] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14).* 1724–1734.

[20] Murk Chohan, Adil Khan, Muhammad Saleem Mahar, Saif Hassan, Abdul Ghafoor, and Mehmood Khan. 2020. Image captioning using deep learning: A systematic literature review. *Int. J. Adv. Comput. Sci. Applic.* 11, 5 (2020).

[21] Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu, and Derry Wijaya. 2020. Informativity in image captions vs. referring expressions. In *Proceedings of the Probability and Meaning Conference (PaM'20).* 104–108.

[22] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19).* 8299–8308.

[23] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10578–10587.

[24] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3298–3308.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 248–255.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* 4171–4186.

[27] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2021. Image captioning as an assistive technology: Lessons learned from VizWiz 2020 challenge. *arXiv preprint arXiv:2012.11696* (2021).

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations.*

[29] Ahmed Elhagry and Karima Kadaoui. 2021. A thorough review on recent deep learning methodologies for image captioning. *arXiv preprint arXiv:2107.13114v1* (2021).

[30] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2022. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22).* 17988–17998.

[31] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19).* 4125–4134.

[32] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 457–468.

[33] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. StyleNet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17).* 955–964.

[34] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17).* 1141–1150.

[35] Lizhao Gao, Bo Wang, and Wenmin Wang. 2018. Image captioning with scene-graph based semantic concepts. In *Proceedings of the 10th International Conference on Machine Learning and Computing.* 225–229.

[36] Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 123–135.

[37] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the International Conference on Machine Learning*. 1243–1252.

[38] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*, Vol. 1. 1440–1448.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning (Adaptive Computation and Machine Learning Series)*. The MIT Press.

[40] Google. 2021. *Open Images Dataset V6 + Extensions*. Google. Retrieved from https://storage.googleapis.com/openimages/web/index.html

[41] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 6837–6845.

[42] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 519–535.

[43] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10323–10332.

[44] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 1969–1978.

[45] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. MSCap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4204–4213.

[46] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. 417–434.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 770–778.

[48] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. 2020. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision (ACCV'20)*. 153–169.

[49] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, Vol. 32, IEEE, 11137–11147.

[50] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. 2001. *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-term Dependencies*. 237–243.

[51] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[52] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* 51, 6, Article 118 (2019).

[53] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3588–3597.

[54] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2022. ExpansionNet v2: Block static expansion in fast end to end training for image captioning. *arXiv preprint arXiv:2208.06551* (2022).

[55] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pretraining for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 17959–17968.

[56] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. 2021. VIVO: Visual vocabulary pre-training for novel object captioning. *Proc. AAAI Conf. Artif. Intell.* 35, 2 (2021), 1575–1583.

[57] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 4700–4708.

[58] Jia-Hong Huang, Ting-Wei Wu, and Marcel Worring. 2021. Contextualized keyword representations for multi-modal retinal image captioning. In *Proceedings of the International Conference on Multimedia Retrieval*. 645–652.

[59] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4634–4643.

[60] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045* (2023).

[61] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 510–526.

[62] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*.

[63] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73.

[64] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.* 128, 7 (2020), 1956–1981.

[65] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. 2019. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953* (2019).

[66] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005* (2022).

[67] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8928–8937.

[68] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[69] Wei Li, Zhaowei Qu, Haiyu Song, Pengjie Wang, and Bo Xue. 2020. The traffic scene understanding and prediction based on image captioning. *IEEE Access* (2020), 1420–1427.

[70] Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Trans. Multim.* 21, 8 (2019), 2117–2130.

[71] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. 121–137.

[72] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[73] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*. 1270–1279.

[74] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. 2022. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 17969–17978.

[75] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS'04)*. 74–81.

[76] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV'14)*. 740–755.

[77] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Kai Lei, and Xu Sun. 2020. Exploring and distilling cross-modal information for image captioning. In *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence*. 5095–5101.

[78] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2023. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506* (2023).

[79] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 873–881.

[80] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. CPTR: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804* (2021).

[81] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 353–369.

[82] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[83] Ren C. Luo, Yu-Ting Hsu, Yu-Cheng Wen, and Huan-Jun Ye. 2019. Visual image caption generation for service robotics and industrial applications. In *Proceedings of the IEEE International Conference on Industrial Cyber Physical Systems (ICPS'19)*. 827–832.

[84] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 5988–5999.

[85] Burak Makav and Volkan Kılıç. 2019. A new image captioning approach for visually impaired people. In *Proceedings of the 11th International Conference on Electrical and Electronics Engineering (ELECO'19)*. 945–949.

[86] Alexander Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating image descriptions with sentiments. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. 3574–3580.

[87] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[88] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).

[89] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2022. GRIT: Faster and better image captioning transformer using dual visual features. In *Proceedings of the European Conference on Computer Vision*. Springer, 167–184.

[90] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc., 4790–4798.

[91] OpenAI. 202. *GPT-4*. Retrieved from https://openai.com/research/gpt-4

[92] OpenAI. 2022. *Introducing ChatGPT*. Retrieved from https://openai.com/blog/chatgpt

[93] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10971–10980.

[94] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 311–318.

[95] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*. 1310–1318.

[96] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proceedings of the 2nd Workshop on Shortcomings in Vision and Language*. 26–36.

[97] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.* 123, 1 (2017), 74–93.

[98] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. 647–664.

[99] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. 8748–8763.

[100] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[101] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR'16)*.

[102] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 91–99.

[103] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1179–1195.

[104] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 4035–4045.

[105] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Trans. Neural Netw.* 20, 1 (2008), 61–80.

[106] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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 (Volume 1: Long Papers)*. 2556–2565.

[107] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 4155–4164.

[108] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 12516–12526.

[109] Shutterstock. 2019. *Stock Images, Photos, Vectors, Video and Music|Shutterstock*. Retrieved from https://www.shutterstock.com/

[110] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. TextCaps: A dataset for image captioning with reading comprehension. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. 742–758.

[111] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15).*

[112] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. 4278–4284.

[113] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[114] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 6612–6621.

[115] Jitendra Vikram Tembhurne, Md Moin Almin, and Tausif Diwan. 2022. Mc-DNN: Fake news detection using multichannel deep neural networks. *Int. J. Semant. Web Inf. Syst.* 18, 1 (2022), 1–20.

[116] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc. 5998–6008.

[117] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[118] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc., 3630–3638.

[119] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[120] Dalin Wang, Daniel Beck, and Trevor Cohn. 2019. On the role of scene graphs in image captioning. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN'19)*. 29–34.

[121] Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan. 2020. Learning visual relationship and context-aware attention for image captioning. *Pattern Recog.* 98 (2020).

[122] Qingzhong Wang and Antoni B. Chan. 2018. CNN+ CNN: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019v1* (2018).

[123] Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022. End-to-end transformer based model for image captioning. *Proc. AAAI Conf. Artif. Intell.* 36, 3 (2022), 2585–2594.

[124] Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. 2018. Scene graph parsing as dependency parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 397–407.

[125] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 203–212.

[126] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. 2021. XGPT: Cross-modal generative pre-training for image captioning. In *Proceedings of the 10th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC'21)*. 786–797.

[127] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 3097–3106.

[128] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37. 2048–2057.

[129] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. 2019. Scene graph captioner: Image captioning based on structural visual representation. *J. Vis. Commun. Image Repres.* 58 (2019), 477–485.

[130] Ning Xu, Hanwang Zhang, An-An Liu, Weizhi Nie, Yuting Su, Jie Nie, and Yongdong Zhang. 2020. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Trans. Multim.* 22, 5 (2020), 1372–1383.

[131] Yang Xu, Li Li, Haiyang Xu, Songfang Huang, Fei Huang, and Jianfei Cai. 2022. Image captioning in the transformer age. *arXiv preprint arXiv:2204.07374* (2022).

[132] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph R-CNN for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 690–706.

[133] Xuewen Yang, Yingru Liu, and Xin Wang. 2022. ReFormer: The relational transformer for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery, 5398–5406.

[134] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10685–10694.

[135] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.

[136] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 711–727.

[137] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 4904–4912.

[138] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 4651–4659.

[139] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Ling.* 2 (2014), 67–78.

[140] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[141] Pengpeng Zeng, Haonan Zhang, Jingkuan Song, and Lianli Gao. 2022. S2 transformer for image captioning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*. 1608–1614.

[142] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 3107–3115.

[143] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5579–5588.

[144] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 6877–6886.

[145] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. 2020. Comprehensive image captioning via scene graph decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. 211–229.

[146] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision—Language pre-training for image captioning and VQA. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. 13041–13049.

[147] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J. Corso. 2017. Watch what you just said: Image captioning with text-conditional attention. In *Proceedings of the Thematic Workshops of ACM Multimedia*. Association for Computing Machinery, 305–313.

[148] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.

[149] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

Received 23 November 2022; revised 11 August 2023; accepted 21 August 2023