Analysis Evolution of Image Caption Techniques: Combining Conventional and Modern Methods for Improvement

Nuha M. Khassaf [1,*] and Nada Hussein M. Ali [2]

[1] Informatics Institute for Postgraduate Studies, Information Technology & Communications University, Baghdad, Iraq
[2] Department of Computer Science, college of science, University of Baghdad, Baghdad, Iraq
Email: phd202230705@iips.edu.iq (N.M.K.); nada.husn@sc.uobaghdad.edu.iq (N.H.M.A.)
*Corresponding author

Abstract—**This study explores the challenges in Artificial Intelligence (AI) systems in generating image captions, a task that requires effective integration of computer vision and natural language processing techniques. A comparative analysis between traditional approaches such as retrieval-based methods and linguistic templates) and modern approaches based on deep learning such as encoder-decoder models, attention mechanisms, and transformers). Theoretical results show that modern models perform better for the accuracy and the ability to generate more complex descriptions, while traditional methods outperform speed and simplicity. The paper proposes a hybrid framework that combines the advantages of both approaches, where conventional methods produce an initial description, which is then contextually, and refined using modern models. Preliminary estimates indicate that this approach could reduce the initial computational cost by up to 20% compared to relying entirely on deep models while maintaining high accuracy. The study recommends further research to develop effective coordination mechanisms between traditional and modern methods and to move to the experimental validation phase of the hybrid model in preparation for its application in environments that require a balance between speed and accuracy, such as real-time computer vision applications.**

Keywords—**Convolutional Neural Networks (CNN), image caption, conventional methods, modern methods, hybrid approach**

I. Introduction

Generating well-structured sentences in image captions requires a deep grammatical and semantic understanding of language based on object detection and recognition, scene type or location, and object attributes and interactions. This field is important and widely used in fields of artificial intelligence, and it focuses largely on understanding images and producing accurate descriptions of them [1]. There are two main types of methods used to generate image captioning: conventional methods and modern methods. Advances in deep learning have allowed computer systems to learn features directly from training data, making them superior to traditional methods in several aspects [2].

Historically, early image captioning research has relied on retrieval strategies and fixed templates. Retrieval-based methods rely on similarity measures to extract appropriate texts from similar images, but they struggle to describe new or unfamiliar scenes. In contrast, template-based methods rely on straightforward steps such as phrase extraction and caption synthesis but are limited in their grammatical diversity and repetition of the same patterns [3]. However, these methods have advantages, such as processing speed and accuracy in simple cases.

On the other hand, neural networks have emerged as the basis for modern methods. These methods are based on the encoder-decoder architecture, where Convolutional Neural Networks (CNNs) understand the content of an image, while Recurrent Neural Networks (RNNs) or transformers decode this content into textual descriptions. These frameworks are enhanced with attention strategies that rely on self-attention to analyze fine details in images [1]. Transformers have become particularly effective in processing complex data sequences and long-term relationships within an image. In addition, reinforcement learning techniques are applied to improve performance in some fine-grained tasks [4]. Despite the success of modern methods, such as deep learning-based detection algorithms like YOLOv4 [5], a research gap remains: the need to balance the simplicity and speed of traditional methods with the flexibility and accuracy of modern models. Effective integration between the two approaches can be achieved to improve the caption generation process, especially in applications that require both accurate and fast descriptions. This study seeks to bridge this gap by exploring the possibility of combining the advantages of the two methods to enhance the accuracy and reliability of the resulting descriptions while maintaining the efficiency and flexibility of performance in handling image diversity. In this context, the study presents a proposed theoretical framework that combines the two approaches.

Manuscript received March 1, 2025; revised March 25, 2025; accepted April 30, 2025; published August 7, 2025.

doi: 10.18178/joig.13.4.406-418

II. Materials and Methods

This study adopts a theoretical and analytical approach to explore the strengths and weaknesses of techniques for generating image captions, combining insights from traditional and modern methods. The proposed hybrid framework was designed based on a comprehensive comparative analysis of previous work, and its effectiveness was evaluated based on a literature review and established standards in the field.

A. Classification and Review of Techniques

The techniques used for caption generation are classified into:

- Traditional methods such as retrieval-based models based on a query set to identify similar images and templates, which rely on predefined linguistic structures. These methods are fast but lack the ability to adapt to complex content.
- Modern methods include deep learning models such as CNN and Long Short-Term Memory (LSTM), transformer models, and advanced visual-linguistic models such as Contrastive Language–Image Pre-training (CLIP) and Bootstrapping Language–Image Pre-training (BLIP). These models are characterized by accuracy and the ability to understand context, but they require high computational resources.

B. Theoretical Integration Strategy

Since this research is theoretical, a set of realistic assumptions, based on what is commonly found in the literature, was adopted to develop this model. These assumptions include:

- The use CNN network such as Xception as an image encoder.
- The use of a two-layer LSTM unit with a storage capacity of 512 units in the linguistic part of the model.
- The reliance on specific linguistic templates such as “there is [object] in [location]” or “the subject is [verb] in [scene].”
- The use of a fusion mechanism based on contextual matching between the caption generated by the traditional model and the potential outputs of the modern model, based on a semantic similarity measure such as Bilingual Evaluation Understudy (BLEU) or Cosine Similarity.

Fig. 1 shows the hybrid approach where the fusion strategy is implemented by calculating the degree of similarity between the descriptions generated by both the traditional and modern methods with the image content, such that each description is given a specific weight based on its consistency with the visual context. The final description is then determined or generated by either selecting the highest-conforming description or combining the strongest elements of both descriptions into a single sentence. This approach achieves an effective balance between the accuracy of modern models and the speed of traditional methods.

These virtual configurations aim to build a logical framework that facilitates analytical evaluation and provides a flexible foundation for future practical experiments. It is also proposed to explore the strategy of early integration of features and traditional feedback to promote deep interaction between visual and linguistic components and improve the quality of generation.

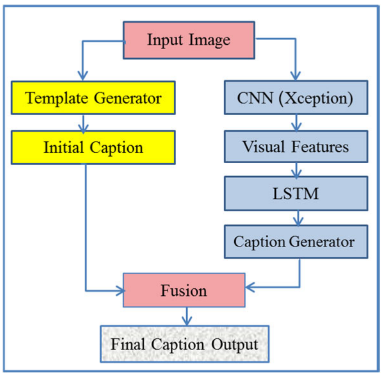

Fig. 1. Hybrid approach, fusion between template method and Convolutional Neural Networks (CNNs)- Long Short-Term Memory (LSTM) model.

C. Benchmark Datasets

To theoretically evaluate the effectiveness of the hybrid model, we reviewed benchmark datasets used in previous studies, including Microsoft Common Objects in Context (MSCOCO). This is one of the most prominent datasets used for training and testing caption models due to the diversity of images and the multiple descriptions per image and Flickr8k/Flickr30k are relatively small datasets but widely used for testing the basic performance of models. The results of published studies using these datasets were used to estimate the theoretical performance of the proposed model.

D. Theoretical Performance Evaluation

Popular performance metrics in the field, such as BLEU, Metric for Evaluation of Translation with Explicit Ordering (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and CIDEr, were used to accurately analyze the quality of the predicted descriptions. The hybrid model is expected to achieve improved results in terms of the balance between accuracy and speed compared to individual methods.

E. Efficiency and Suitability Considerations

Although the hybrid approach requires implementing two methods (traditional and modern), which may increase the overall computational load, using the traditional method in the initial generation phase contributes to reduced response time compared to relying solely on deep models. Thus, the system achieves a balance between efficiency and effectiveness, making it suitable for environments that require fast performance without sacrificing accuracy. The integration mechanism also allows for flexibility in customization, as the level of computational complexity can be controlled according to the application requirements.

F. Future Prospects

This methodology provides a comprehensive conceptual foundation that paves the way for future experimental validation and can be built upon to develop more efficient and flexible hybrid models in computer vision and artificial intelligence applications. The potential for early integration of traditional and modern methods within the hybrid model could also be explored in the future, with the goal of enhancing interaction between components and achieving higher performance in complex environments.

III. Conventional Methods Captioning

Early efforts to teach computers to understand the visual content of images were initiated by using retrieval and template-based methods to generate annotations. Although simple, these methods are limited in their capabilities, as they are unable to generate innovative captions or adapt text to fit the unique details of each image [6]:

A. Retrieval-based Method

Image captioning is a retrieval job in this type of approach. This technique finds the similarity metric score for classification or maps texts and images into a shared vector space. Comparing the input image to the query set to see whether any images are similar [7]. For the matching candidates of the recovered images, the caption with the best explanation is selected from the chosen image caption [3]. One or more sentences from the corpus, or a combination of both, can be used to generate captions, presuming that the supplied image is comparable to one already within a database where the computer immediately uses the annotation of the retrieved image to describe this image [7].

Hodosh *et al.* (2013) [8] applied the Kernel Canonical Correlational Analysis (KCCA) to sentence-based picture description On the Flickr 8k dataset, where 94.7% of the items that passed the criteria had an expert score of 2.7 or higher; KCCA achieved much better outcomes than nearest neighbour-based methods.

Mason and Charniak (2014) [9] proposed a non-parametric an intensity estimation method that involves defining a visual similarity feature space and then formulating intensity estimation problem to model words used to describe visually similar images, outperforming the Scene Attributes and Collective systems in terms of relevance by 48% and 34%, respectively.

Devlin *et al.* (2015) [10] proposed a method that finds k-Nearest Neighbour (k-NN) images from training datasets and comes with a suitable caption. The method achieved 27.6% by k-NN, way better than humans.

Ordonez *et al.* (2016) [11] demonstrated two forms of text retrieval: one that retrieves the complete image description and another that retrieves specific items or scenes based on their visual and geometric similarities. (BLEU = 0.1260).

B. Template-Based Method

This method generates descriptive phrases by specifying a strict sentence structure in grammatical form and filling it with predicted nouns, verbs, and scenarios [12]. Template-based techniques ensure sentence grammar. Many template-based translation systems extract one or more words from the image. Subjects, predicates, and prepositions are attached to the descriptions [13].

Kulkarni *et al.* (2013) [13] proposed using Conditional Random Fields (CRF) to identify the appropriate words to describe an image from a database of visual illustrations (BLEU = 0.21).

Elliott and Keller (2013) [14] demonstrated how to describe an image through dependencies between its annotated parts. This visual dependency representation encodes which regions are associated with the image and is used to infer the action or event depicted. (BLEU-1 = 45.4, BLEU-2 = 16.1, BLEU-3 = 6.4, BLEU-4 = 2.70).

Lebret *et al.* (2015) [15] proposed a model that can infer different phrases from image samples. The predicted phrases are then used through a statistical language model to generate sentences. The model achieved (BLEU-1 = 0.60, BLEU-2 = 0.37, BLEU-3 = 0.22, BLEU-4 = 0.14) on Flickr30k dataset.

Xu *et al.* (2015) [16] used embedded texts in continuous vector space using dependency trees where visual meaning and word order can be preserved, achieving a rank mean of 224.1 and finding phrase-based sentences more descriptive than word-based ones.

C. Discussion

Conventional methods, such as retrieval and template methods, face challenges in representing semantic features due to the "semantic gap" problem, which makes them non-generalizable and limited in flexibility [9]. The retrieved captions may not accurately match complex images [8], and the descriptions generated by template-based methods may be static and monotonous [13]. However, these methods are fast and simple, requiring less execution time and not requiring complex model training. This makes them suitable for applications that require fast responses, making them attractive in certain contexts, despite their limitations in handling complex data [15].

IV. Modern Captioning Methods

Most modern techniques used to generate a caption for images are based on deep learning and transformers, where Fig. 2 illustrates typical architectures that rely on deep learning, such as CNN-RNN and Transformer-based models, which extract visual features and generate textual descriptions accordingly. Many researchers have proposed different effective optimization methods; each has different emphases and divides into multiple subcategories according to enhanced focus [17].

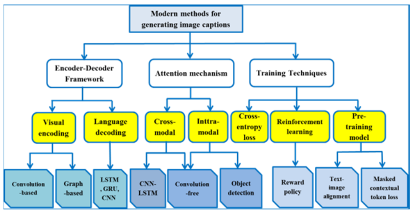

Fig. 2. Taxonomy of Modern methods for image caption generation.

A. Frameworks of Encoder-Decoder

Deep learning captioning models use encoder-decoder architectures to handle variable-length sequence inputs and outputs, making them suited for sequence-to-sequence challenges such as machine translation and image caption generation, and include the following:

1) Visual encoding

This encoding aims to extract visual features from an image, such as colors, shapes, and patterns, and then convert these features into a form the model can understand and use to generate descriptive sentences that clearly explain the image. This involves understanding the relationship between different elements in the image and generating a description that clearly expresses them. Two main categories of visual representation learning are convolutional neural network-based representation learning and graph-based representation learning [1]:

a) Convolutional-based

In this context, the use of Convolutional Neural Networks (CNNs) refers to the extraction and representation of visual features from images to support analysis and annotation processes. Fig. 3 illustrates the architecture of a model based on two types of feature representation: grid feature representation, where the image is divided into a regular grid to extract features from each part, and region feature representation, where regions of interest within the image are identified and their features are extracted in a customized manner. Visual features are extracted across multiple layers of the convolutional neural network and subsequently used as inputs for annotation generation modules, contributing to an accurate description of image content [18, 19].

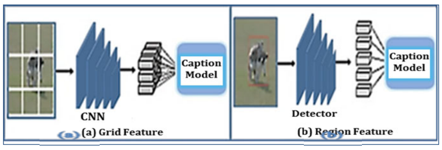

Fig. 3. CNN-based visual encoding strategies: (a) Grid-based approach; (b) Feature-based approach.

Faiyaz *et al.* (2021) [20] used a pre-trained ResNet-50 model image encoder to extract region-based visual characteristics and a single-dimensional CNN to encode sequencing in image captioning. The study achieved (BLEU-1 = 0.651, CIDEr = 0.572, METEOR = 0.297, ROUGE = 0.434).

Shinde *et al.* (2021) [21] used VGG16 and LSTM Models. The model takes an image as input and by analysing the image, it detects objects in the image and creates a suitable caption for them; the model achieved 92.7% top-5 test accuracy.

Zhang *et al.* (2021) [22] used pre-trained CNN to get global network characteristics and learn explicit representations of high-level features to improve captions. (BLEU-1 = 82.1, BLEU-2 = 67.0, BLEU-3 = 52.2, BLEU-4 = 40.0).

Datta *et al.* (2019) [23] proposed a method for phrase grounding using the weak supervision available from pairs of images and corresponding captions. Their contribution lies in the design of the local pooling module, which plays a major role in the tight coupling; the model achieved (R@ = 56.6, R@5 = 84.9, and R@10 = 92.8) on the COCO dataset.

Chen *et al.* (2020) [24] introduced the visual idea detector and LSTM caption generator, which can gain more visual and semantic knowledge from out-of-band images and text. (BLEU-1 = 68.7, BLEU-2 = 50.7, BLEU-3 = 36.6, BLEU-4 = 26.1).

Yang *et al.* (2021) introduced a technique for incorporating spatial coherence of objects into a model of image caption. For each two overlapping objects, concatenates their initial visual features to generate two directional pairwise features and learns weights that optimize these pairwise features, resulting in the model (BLEU-1 = 57,63, BLEU-2 = 35.58, BLEU-3 = 23.63, BLEU-4 = 14.14) [25].

b) Graph-based

Some recent studies have used image region-based graphs to improve image representation by adding semantic and geographic relationships between regions, enhancing the encoding of object interactions within the image [26].

Graph Convolutional Neural networks (GCNs) are a type of multilayer neural network that operates directly on data represented as graphs, with information transmitted via edges between nodes. In the context of caption generation, CNNs are used to extract visual features, while GCNs are used to analyse semantic relationships between objects in the image. Fig. 4 shows both CNNs and GCNs are combined into a single architecture. The CNN is used to extract initial features, and these features are then passed to the GCN to construct a semantic representation that improves the quality of the generated caption by understanding the relationships between objects [27].

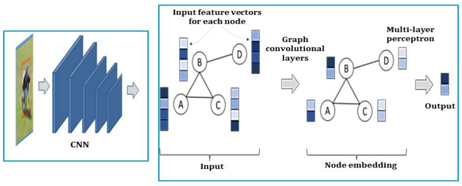
CNN
Input feature vectors for each node
B
D
A
C
Graph convolutional layers
Multi-layer perceptron
B
D
A
C
Output
Input
Node embedding

Fig. 4. Combination of Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN) for image caption generation.

Yao *et al.* (2018) [28] introduced a design for exploring inter-object connections for image captioning using GCN and LSTM that integrates semantic and spatial object relationships. CIDEr-D and SPICE scores had boosted to 128.7 % and 22.1%.

Yang *et al.* (2020) [29] used Scene Graph Auto Encoder (SGAE) to integrate language inductive bias into an encoder-decoder image annotation framework for more human-like captioning; their method achieved (BLEU-1 = 60.8, BLEU-4 = 17.1)

Zhang *et al.* (2021) [30] used Consensus Graph Representation Learning (CGRL) to incorporate a consensus representation into the grounded captioning pipeline by aligning the visual graph with the linguistic graph, which considers nodes and edges. Their method achieved BLEU-1 = 72.9 and BLEU-4 = 28.3.

Nguyen *et al.* (2021) [31] proposed a method that used only scene graph labels to perform competitive image annotation. The idea is to bridge the semantic gap between two scene graphs, derived from the input image and its caption; the method achieved (BLEU-1 = 75.0, BLEU-4 = 32.6).

c) Discussion

The two main types of visual representation learning for caption generation are CNN-based and graph-based approaches, where Table I compares the two approaches in terms of description, highlighting their strengths and weaknesses [18]. In CNN-based models, convolutional networks such as VGGNet and ResNet are used to extract high-level visual features from images [20]. This enables the model to effectively represent objects, although its ability to distinguish between important and unimportant regions of an image is sometimes limited [21]. Graph-based models, such as GCNs, enhance representation by understanding the semantic and structural relationships between objects, but they require complex design to accurately represent spatial and semantic relationships [26]. For example, GCNs play a role in effectively enhancing the semantic description of objects, leading to improved performance in annotation tasks [27]. Using graphs to represent the relationships between different elements within an image allows for a deeper understanding of the semantic structure of the image, which contributes to generating more accurate annotations [29, 30].

TABLE I. COMPARISON BETWEEN CNN-BASED AND GRAPH-BASED VISUAL ENCODING METHODS

Method	Description	Strengths	Weaknesses
Convolutional-based	CNNs are used to extract visual features, models such as VGGNet and ResNet	Extract high-level semantic features (global and local)	Difficulty distinguishing between important and unimportant region
Graph-based	GCNs are used to represent relationships between objects within an image	Captures semantic and spatial relationships between objects, more coherent and context-aware caption generation	Requires sophisticated graph construction and training strategies

2) Language decoding

The decoder's objective to predict the probability of occurrence of a given sequence of words in a phrase treats text generation as a random process. Methods of decoders are distinguished based on the language model that uses [32]:

a) Decoder methods

- LSTM is currently the most popular language model in image captions. LSTMs can learn more complex and long-term patterns thanks to their special structure that allows them to store information for long periods and control the flow of information effectively [33].
- Gated Recurrent Unit (GRU) is simple compared to LSTM and has limited memory. In some cases, as an alternative to LSTM because it is less complex and has fewer parameters.
- CNNs are usually used to extract features from images. However, some research also uses CNNs as a decoding model. Embedding vectors of words and image features are fed into the model (such as an LSTM) to teach it what the image is about. This means that the model learns to associate image features with the correct text patterns to generate descriptive sentences that match image content [34].

Aneja *et al.* (2018) [33] proposed the convolutional captioning of images method that performed similarly to the LSTM baseline on the MSCOCO dataset with a quicker training time per number of parameters and scores (BLEU-1 = 0.725, BLEU-2 = 0.555, BLEU-3 = 0.41, BLEU-4 = 0.299).

Wang *et al.* (2018) [34] suggested a system that solely uses CNNs to produce captions; using parallel processing, the model outperformed the NIC by a factor of three during training and attained scores (BLEU-1 = 0.350, BLEU-2 = 0.194, BLEU-3 = 0.107, BLEU-4 = 0.059).

Wu *et al.* (2019) [35] suggested a recall network for image captions. The network uses Grid LSTM to selectively include visual features and recall image contents while generating each phrase.

Khan *et al.* (2022) [36] used more than one pre-trained CNN as encoding, a language model GRU used as decoding for construction of the descriptive phrase, and merged Bahdanau attention with GRU to improve the results, and the attained scores (BLEU-1 = 0.78, BLEU-2 = 0.57, BLEU-3 = 0.44, BLEU-4 = 0.36) on the MSCOCO dataset.

b) Discussion

Popular decoding models such as LSTM and GRU are among the most widely used models for generating descriptive texts from images. LSTMs provide the ability to learn long-term patterns in sequential data, making them ideal for generating texts that require a comprehensive understanding of the context. LSTMs have been used in an image caption retrieval network model using Grid LSTM to selectively include visual features [35]. On the other hand, GRUs are a simpler and less complex option than LSTMs, making them suitable when resources are limited or when the increased complexity of LSTMs is not necessary when combined with attention mechanisms, which results in good scores [36]. CNNs are used primarily to extract visual features from images and combine them with other models such as LSTMs to generate texts. CNNs were used in parallel with LSTMs to achieve excellent results with faster training time [33]. However, some researchers suggest using CNNs solely as a decoding medium, which led to improvements in training speed [34].

B. Attention Mechanism

Encoder and decoder models are commonly used with attention mechanisms in image translation and generating descriptive labels to achieve a deep and accurate understanding of an image visual and semantic attention to coherent image regions and interest elements is required [37]. Attention mechanisms, including intra-attention and cross-attention, help identify the most important parts of the input data, such as images or texts, that the model needs to focus on. Intra-attention analyses the internal relationships within the same data type. While cross-attention enhances the information integration between images and texts, improving the accuracy of image labels by providing accurate descriptions based on an integrated analysis of visual and linguistic elements [38].

1) Cross-modal attention

Cross-media attention focuses on integrating information from different media, such as text and images. Features from one medium are used to improve understanding of another medium. Cross-modal attention can be included in multiple models to improve data processing [39]:

- CNN-LSTM: The CNN is used to extract visual features from images, while LSTM is used to process texts. Cross-modal attention can be combined with this model to increase the focus on specific features of the image and text, which improves the quality of image description; models include the Attend and Tell Model, One-layer LSTM Attention, and Two-layer LSTM with Dual Attention.
- Convolution-free: Transformers that rely on self-attention are used to process features without needing traditional convolution layers. Cross-modal attention can also be included in these models to improve information integration between images and texts, which enhances the accuracy of the description.

Cao *et al.* (2020) [40] presented a method based on learning the interaction between images and descriptions using a parallel, convolutional-free attention network, attention weights for these images and words are determined based on their mutual relationship. This method achieved Recall = 0.4960 and NDCG = 0.3829.

Liu *et al.* (2020) [41] used an efficient way to explore and distill source information across media using a transformer. Global distillation methods learn to capture clusters of salient regions and features while exploring fine-grained spatial and relational representations. This method achieved CIDEr = 129.3.

Zhang *et al.* (2021) [42] suggested a method for cross-media semantic content mapping to link images and captions. The model uses a joint attention network to query image-text pairs and determine the dependence of words on visual content. These relationships are integrated into the LSTM network for sentiment analysis, this method achieved an accuracy = 0.806.

Pourkeshavarz *et al.* (2023) [43] presented an attention network that stacks cross-modal features for consolidation using a compounding function in a multi-step reasoning process. The model also includes using CNN to extract visual features and an LSTM network to generate text captions based on these combined features. Experimental results showed that model achieved BLEU-1 = 71.2, and BLEU-4 = 27.9.

2) Intra-modal attention

Intra-modal attention focuses on processing information within only one medium, whether that medium is an image or text [44]:

- Object Detection: R-CNN, YOLO etc. which based on Convolutional Neural Networks (CNNs) to detect objects in image. While these methods mostly rely on convolutional networks, they can benefit from adding attention mechanisms to allow better focusing what is truly important in the image to enhance recognition accuracy [45].
- Convolution-free: Techniques such as transformers that rely on self-attention are used to process features within the same medium, whether it is text or an image. These methods allow focusing on the relationships between different parts of the same medium, improving the efficiency of data processing without relying on traditional convolutional networks. Transformers provide a powerful alternative to convolutional networks in processing data within a single medium [46].

Zhu *et al.* (2018) [47] replaced LSTM with transformer decoding using Stacked Self-Attention, which solves the cross-time sequence problem that traditional models suffer from. This method enhances intra-model attention by improving the focus on the relationships between different parts of the same medium, which contributes to improving the efficiency of data processing. This method achieved (BLEU-1 = 72.9, BLEU-4 = 33.1).

Yu *et al.* (2019) [48] presented a Multimodal Transformer (MT) framework that integrates a visual encoder to generate visual representations via self-attention, and a decoder to convert these features into textual captions. This model enhances intra-modal interaction by optimizing image and text representations internally, and enhances cross-modal interaction by integrating visual and textual information; the method achieved (BLEU-1 = 81.7 and BLEU-4 = 40.4).

Herdad *et al.* (2019) [49] used an object relation transformer to improve the interaction between objects within an image. The method relies on incorporating information about the spatial relationships between the specified objects using geometric attention, which makes it free of convolutions and enhances the internal interaction between objects, a method achieved (BLEU-1 = 80.5, BLEU-4 = 38.6).

Guo *et al.* (2020) [50] proposed using Natural Self-Attention (NSA) to reduce the effect of the internal variable transformation. Geometry-aware Self-Attention (GSA) is used to explicitly and dynamically compute the geometric bias between objects to improve image understanding. The achieved results (BLEU-1 = 80.8 and BLEU-4 = 38.8).

Liu *et al.* (2021) [51] used a transformer that processes concatenated raw images and applies global context models at each encoding layer, removing convolutions and redundancies, this method achieved (BLEU-1 = 81.8, BLEU-4 = 39.5).

Sundaramoorthy *et al.* (2021) [52] used an end-to-end transformer model, where shallow layers rely on multiple attention heads to exploit local and global contexts, which CNN encoders cannot achieve. The model achieved a superior Levenshtein distance of 6.95 on average compared to ResNet/LSTM with attention of 7.49.

Iwamura *et al.* (2021) [53] used motion features and object detection to improve the generation of annotations. Feature extraction from object regions was used instead of all motion features to increase accuracy. The (BLEU-1 = 75.9, BLEU-2 = 59.9, BLEU-3 = 46.0 and BLEU-4 = 35.2).

3) Discussion

Both cross-pattern and intra-pattern attention contribute to improving the quality and accuracy of image description generation. Cross-pattern attention enhances the model's ability to focus on important elements in both images and text, leading to improved integration and description accuracy when using CNNs for feature extraction with LSTM models for text processing [42, 43]. With the development of self-attention techniques, transformers have emerged as an effective alternative that allows for more efficient integration of textual and visual information without relying on traditional convolutional networks [40, 41].

TABLE II. COMPARISON BETWEEN CROSS-MODAL AND INTRA-MODAL ATTENTION

Type of Attention	Description	Strengths	Weaknesses
Intra-Modal Attention	Processing information in same medium , such as R-CNN, YOLO, Self-Attention, Vision Transformers	Enhances accuracy by focusing on key image regions	Requires powerful techniques for identifying important parts, less efficient at handling complex textual
Cross-Modal Attention	Linking information between different media(images and texts), Attend and Tell, One-layer LSTM Attention, Two-layer LSTM with Dual Attention, Transformers	Improve media interaction, generate accurate descriptions, handle complex information	Complexity of models, need for high computational capabilities, require large data to train

Intra-pattern attention enhances focus on essential parts within the same medium. In image processing, techniques such as R-CNN and YOLO rely on convolutional networks for object detection, while transformers allow for direct processing of visual features [47–50]. In texts, intra-pattern attention is used to improve understanding of relationships between different textual components [51, 52]. Table II compares the two types of attention, explaining their mechanisms of action and the key strengths and weaknesses associated with each.

C. Training Techniques

Various training techniques are used, such as cross-entropy loss to improve the probabilities of individual words, reinforcement learning to improve the quality of entire sequences through specific rewards, and pre-training models such as Bidirectional Encoder Representation from Transformers (BERT) and pre-trained Generative Transformers (GPT) to provide robust linguistic representations [54].

Recently, advanced language-vision models such as CLIP, BLIP, and Flamingo have emerged, relying on large multimodal datasets, enabling a deeper understanding of the relationship between images and text. These models offer significant improvements over traditional pre-training, combining image and text representations into a common space, helping produce more accurate and context-rich captions. These techniques improve the performance of image labelling models and increase the accuracy of text predictions [55]. Radial search is used in the prediction phase to improve the quality of the resulting text based on the probabilities and predictions produced by these techniques [56].

1) Cross-entropy loss

Cross-entropy loss is used during the training model to predict the next words in the sequence based on the previous words. The goal is to minimize the gap between the predictions and the actual sequence of words. During the training phase, the models improve their performance using cross-entropy loss. However, in the prediction (decoding) phase, beam search is used to improve the accuracy of the generated sequences based on the learned probabilities. However, this can lead to an accumulation of errors in the verbal sequence, as any error in a particular word can affect the prediction of subsequent words.

During the training process, the model aims to minimize the difference between real and predicted words by minimizing the negative logarithm of the probability of the correct word, using a cross-entropy loss function, which is widely used in sequence prediction tasks, such as generating image captions. This is represented mathematically in Eq. (1) [56].

$$L_{XE}(\theta) = -\sum_{i=1}^{n} log(P(y_i \mid y_{1:i-1}, X)) \quad (1)$$

where X is visual encoding, Y_i grounding-truth word at the time i, $y_{1:i-1}$ preceding grounding-truth words, P probability distribution.

Li *et al.* (2020) [57] investigated the effect of VGG16 encoder modification on image captioning tasks using cross-entropy loss. Their method achieved BLEU-1 = 0.919, reflecting a significant improvement in the quality of the resulting captions.

Maru *et al.* (2021) [58] compared the VGG16 and InceptionV3 architectures and studied the effect of cross-entropy loss and Kullback-Leibler divergence (KL) variation on model training, where the InceptionV3 achieved BLEU-1 = 0.93, while the VGG16 achieved BLEU-1 = 0.919, highlighting the effectiveness of cross-entropy loss in improving the performance of models.

2) Reinforcement Learning (RL)

RL is used to overcome cross-entropy limitations by evaluating the entire sequence of words rather than each word individually. The model learns how to optimize its rewards across the sequence, which helps improve the quality of the generated text. When using RL, an action selection strategy is implemented at each time step by predicting the next word to maximize rewards across the entire sequence. This improves the model's ability to generate sequences that are more accurate and reduces error accumulation better than using cross entropy alone [59]. During prediction, beam search can be used along with reinforcement learning strategies to improve the quality of the generated sequences. Reinforcement learning helps the model improve its word selection policy by evaluating the generated labels based on a given reward, while beam search allows for retaining multiple candidate probabilities at each time step to enhance the quality of the generation. The loss gradient generated by algorithms such as beam search and greedy decoding is calculated using Eq. (2) [56].

$$\nabla_\theta \, \mathrm{L}\,(\theta) = -\frac{1}{K}\sum_{i=1}^{K}((r(w^i)-b)\,\nabla_\theta \, log \;\; p(w^i)) \qquad (2)$$

where w^i is i-th phrase in beam or sampled collection, reward operation is denoted by r, and b baseline is computed using reward obtained from greedy decoding.

Liu *et al.* (2018) [60] proposed a multilevel reward function that combined vision and language rewards to guide policies, which helped improve the model performance and achieved (BLEU-4 = 0.330).

Seo *et al.* (2020) [61] used a policy gradient method to improve people's evaluations as rewards in a reinforcement learning environment, where policy gradients were estimated based on annotations and achieved an average score of 68.42 ±0.61%.

Honda *et al.* (2023) [62] proposed lightweight fine-tuning methods to address the bottleneck in reinforcement learning models, resulting in improved recognition accuracy with only minor adjustments, achieving smooth matching (RefCLIPS = 81.5).

3) Pre-training model

Recently, Visual Language Pre-training (VLP) approaches whose critical goal is the joint learning of visual and textual features through transformer-based architecture flourished, showing new advancements in different vision language jobs. Two aims for pre-training models are as follows [63]:

- Text-image alignment: Text-image alignment is a key focus of modern pre-training models. It aims to enable models to understand the contextual relationship between visual and textual content, contributing to the generation of accurate and informative descriptions. This field has seen significant development thanks to advanced multi-modal models.

In this context, recent models such as CLIP, BLIP, Flamingo, and GPT-V have made significant progress in text-image alignment. CLIP relies on learning from massive amounts of text and image data, enabling it to generate rich representations that link words and visual entities without the need for extensive supervision. BLIP expands the capabilities of CLIP by incorporating text paraphrasing techniques and improving multimodal context capture. Flamingo, on the other hand, is a flexible model that combines few-shot learning with the ability to interact with images and text, making it suitable for multiple tasks in description generation and visual interaction. Finally, GPT-V enhances the ability of generative transformers to generate rich descriptive texts based on in-depth analysis of images and associated texts. These models have contributed to improving the accuracy of annotation models, but they require significant computational resources.

Zhou *et al.* (2020) [64] presented a unified VLP model: (1) Finely tuned for either vision-language creation or comprehension (visual question answering) tasks, (2) uses the shared multi-layer transformer network for each one (encoding and decoding). The model achieved (BLEU-4 = 36.5).

Mokady *et al.* (2021) [65] proposed the CLIP representation as a "prefix" that is fed to a language model like GPT-2 to generate text captions for images. This is achieved through a simple convolutional network that connects the CLIP outputs to the language model inputs, as only the convolutional network is trained while CLIP and GPT-2 remain unmodified, results of method (BLEU-4 = 32.15, METEOR = 27.1, CIDEr = 108.35) on COCO dataset.

Li *et al.* (2022) [66] suggested BLIP, a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a caption generates synthetic captions and a filter removes the noisy ones. This method achieved (+2.8% in CIDEr).

Alayrac *et al.* (2022) [67] proposed model is the Flamingo model, which combines pre-trained models for vision and text and relies on few-shot learning. The model is trained on multi-modal data to enable it to perform tasks such as generating captions and answering visual questions with flexibility and high accuracy.

Wang *et al.* (2023) [63] used Masked Language Modelling (MLM) and cross-distillation techniques to generate smooth labels aimed at improving the robustness of the model. They also used Image-Text Matching (ITM) techniques with a linguistic encoder to incorporate hard negatives that depend on the language input context. Model achieved (BLEU-4 = 41.0).

Jin *et al.* (2023) [68] proposed a methodology applicable to tasks such as visual question answering, annotation, and grounding with little training. This methodology relies on text alignment with the topic to help gain an understanding of the topic and locate it. The model achieved (R10 = 76.0).

- Chen *et al.* (2024) [69] proposed GPT-4V model uses visual and camera recognition to integrate text and visual context, enabling feature generation using both visual and textual context. The model predicts leading labels based on visual and textual contributions without requiring specialized training. Masked contextual token loss: Pre-training aims to use a masked contextual token loss technique, similar to BERT, where textual and visual tokens are randomly hidden. To learn a joint representation, the model is asked to predict the hidden inputs based on their surrounding context. This approach enhances the model's ability to understand contexts and interactions between text and images, leading to improved performance on various visual language tasks [70].

Gao *et al.* (2019) [70] proposed a reformulation of feature function to estimate the feature of each token without the need for an additional parameter. The revised multi-step feature approach increased the absolute value of the mean, reducing the variance. The method achieved (BLEU-1 = 77.6 and BLEU-4 = 34.8).

Ren *et al.* (2022) [71] used a mask-guided transformer with topic tokens and a multi-head attention mechanism to capture scene features and understand relationships objects. The topic tokens served as a pre-reference in the decoder, enabling it to focus on global information. The method achieved (BLEU-1 = 80.40, BLEU-4 = 54.12).

Some recent research has also begun to explore the potential of diffusion models in generating text from images.

Fatemeh *et al.* (2024) [72] provided a comprehensive review, noting that previous literature has neglected the capabilities of these models in generating accurate captions. The study reviewed the basic principles of diffusion models, routing and conditioning techniques, and provided a classification of recent approaches, emphasizing that these models represent a promising approach to text-image alignment, especially when combined with semantic adaptation techniques.

4) Discussion

Improving models requires a sequential and thoughtful approach to achieve better performance; therefore training techniques are divided into the following stages:

- Cross-Entropy Loss: Used to reduce the gap between model predictions and the actual word sequence, which strengthens linguistic understanding and improves performance [57–58].
- Reinforcement Learning: Reinforcement learning overcomes the limitations of cross-entropy loss by evaluating the entire word sequence. That helps improve the quality of generated texts and reduce error accumulation [59] through reward and model tuning strategies [60–62].
- Pre-training: Pre-training focus on jointly learning visual and textual features using transformers, which enhances the model's ability to better understand context [63, 64, 68]. Text-image alignment plays a pivotal role in this context. Models such as CLIP, BLIP, and Flamingo have demonstrated outstanding performance by relying on efficient attention mechanisms and large multimodal databases [65–67, 69]. Despite challenges, the need for high computational resources and the risk of generating inaccurate descriptions. Masked contextual token loss is used within this framework, where text or visual tokens are randomly masked to train the model to predict them, enhancing deeper understanding of text-image interactions [70, 71].
- Diffusion Models: These models represent a promising approach, contributing to improved text-image alignment by generating accurate captions using guidance and adaptation techniques [72].

V. Hybrid of Conventional and Modern Methods

Unlike previous studies that relied either on traditional methods alone, which often fail to understand complex images, or on state-of-the-art methods alone, which require massive computational resources and can produce inaccurate descriptions in some cases, this research proposes a theoretical hybrid approach based on fusion between two methods,were the approach relies on an interactive and complementary fusion of the results of traditional methods, which provide a quick initial description using templates or retrieval, with the outputs of modern deep learning-based models, such as CNN and LSTM, resulting in a more accurate and balanced final description.

A. Comparison with Traditional and Modern Methods

Several conventional experiments, like the one conducted by J. Kulkarni *et al.* (2013) [13], depended on the use of preset linguistic templates to generate captions. These techniques worked well for basic images but were unable to adjust to complex ones, which led to poor handling of visual diversity and restricted generalization.On the other hand, more recent research has depended on deep models to produce descriptions. For example, Sundaramurthy *et al.* (2021) [52] used Convolutional Neural Networks (CNNs) in their study. Even though these models produced excellent results, they were computationally demanding and occasionally had trouble describing the image accurately because of their poor contextual awareness.

The Table III offers a thorough comparison that emphasizes the benefits and drawbacks of each of these conventional, modern, and suggested hybrid approaches.

B. Support for the Hybrid Trend in the Current Literature

Recent research published at the CVPR 2023 conference has emphasized the importance of combining traditional analysis with deep processing, which promotes the trend toward hybrid models in the fields of computer vision and caption generation:

TABLE III. Comparison between Traditional, Modern and Hybrid Methods

Advantage	Traditional methods	Modern methods	Proposed hybrid approach
Accuracy	Weak with complex images	High performance but sometimes error-prone	High and balanced
Resource Consumption	Very low	Very high	Medium (optimizing performance versus resources)
Flexibility	Inflexible with new content	Flexible but requires massive training data	Flexible and can handle different types of images
Speed of Implementation	Fast but inaccurate	Slow but high accuracy	Balanced between speed and accuracy
Adaptability	Very limited	Requires extensive training	Able to adapt without requiring massive data

Kuo & Kira (2023) [73] proposed Hierarchical Aggregation of Augmented Views for Image Captioning (HAAV), which contributed to improving understanding of visual context and enhancing the accuracy of the resulting descriptions.

Zeng *et al.* (2023) [74] presented a framework for generating multiple, Controllable Zero-shot Image Captioning by Sampling-Based Polishing (ConZIC), enabling the generation of rich and diverse captions without the need for additional training.

Zeng *et al.* (2024) [75] suggested Memory-Augmented Zero-shot Image Captioning (MeaCap), used an external memory mechanism to retrieve relevant linguistic concepts from texts, enabling the model to generate accurate captions even in zero-shot situations, i.e., without direct training on image data.

Some recent research, has also contributed to supporting the principles behind hybrid models, indirectly dedicated to caption generation. For example, Yamini *et al.* [76] presented a model based on a custom transformer (mT5) for generating abstract summaries in Sorani Kurdish, supported by a manually collected and annotated dataset. Although the study focused on textual summaries, it clearly demonstrates the importance of hybrid models in resource-poor languages by combining transformer structures with human data evaluation. This supports the trend toward integrating deep learning and symbolic processing methods in multiple contexts.

In the same vein, Fatemeh *et al.* [77] proposed a flexible model for integrating multi-view representations while preserving their diversity, enhancing the models' ability to handle diverse and heterogeneous information. This idea can be used in the design of hybrid caption generation models, as preserving the distinctness of representations for both traditional and modern paths may contribute to generating more accurate and comprehensive descriptions.

Together, these studies suggest that the move toward a systematic integration of symbolic (traditional) and deep (modern) processing perspectives reflects an advanced research trend, contributing to enhanced generative accuracy and adaptability to different contexts.

C. Applying a Hybrid Approach Based on Fusion

Based on the hypothetical methodology, a theoretical practical implementation of the hybrid approach is proposed, combining the outputs of a traditional template-based model with the outputs of a modern deep learning-based model as show in Fig. 5. This implementation aims to provide a comprehensive vision of how fusion is implemented at the description level, while achieving a balance between generation accuracy and execution speed.

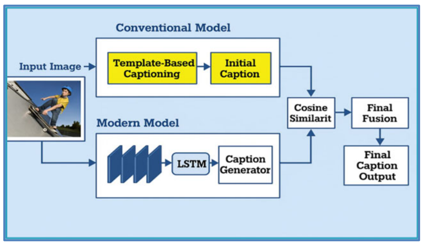

Fig. 5. Diagram of the proposed hybrid model for image captioning.

The theoretical implementation consists of a series of integrated stages, which can be described as follows:

1) Initial generation stage using the traditional model

The image is passed to a module based on the specified linguistic templates (e.g., "[object] is located at [location]"), which produces a rapid initial description reflecting the visual elements clearly visible in the image, such as "a boy is riding a scooter on the street".

2) Extracting features and generating contextual description using CNN-LSTM

The same image is analyzed using the Xception model to extract a high-level semantic visual representation, which is then sent to a two-layer LSTM module to generate a context-based, modern description, such as "a boy wearing an orange helmet riding a scooter".

3) Theoretical fusion mechanism

To determine the final image description, a fusion mechanism based on a semantic similarity measure, such as BLEU or Cosine Similarity, is applied according to the following steps:

Similarity Score Calculation: The semantic consistency between the description generated by the template and the description generated by the neural model is measured.

Relationship Analysis with Image Content: The extent to which each description matches the image details is evaluated to ensure its relevance to the actual visual context.

- Relative Weighting: Each description is weighted based on its semantic similarity scores and the extent to which it expresses the image content, without adhering to a predetermined ratio, with the aim of achieving the highest consistency between the text and the image.

4) Selection or synthesis of the final description

Based on the results of the analysis phase, one of the following two paths is followed:

- Selection: The description that demonstrates the highest degree of consistency with the image and the textual context is selected.
- Synthesis: If the descriptions are close in quality, the most important semantic elements from both descriptions are combined to create a richer, more accurate unified sentence, such as: "a boy wearing an orange helmet riding a scooter on the street".

5) Expected final outcome

The system is expected to produce a final description that achieves the following:

- Higher linguistic accuracy thanks to the use of an LSTM model to generate context-based text.
- Initial speed in visual comprehension by leveraging the traditional template.
- Improved semantic consistency thanks to the application of an effective similarity measure between text and image.

This scenario is based on assumptions that can be verified in a future practical application and aims to develop a flexible model that combines traditional comprehension with deep image representation. It is also proposed to explore more advanced integrations in the future, such as integrating features at a deeper level or implementing iterative interaction mechanisms between traditional and modern components to enhance performance.

D. Benefits of a Hybrid Fusion-Based Approach

Based on the proposed theoretical analysis, the hybrid integration-based approach offers several key benefits:

Interpretive and Analytical Fusion: Combining traditional template rules with deeply extracted features provides a linguistic foundation supported by comprehensive visual understanding.

- Improved Accuracy: Fusion reduces the likelihood of generating generic or illogical descriptions.
- Higher Efficiency: A theoretical review demonstrated that this approach improves computational efficiency by approximately 20% in the initial generation phase, compared to relying entirely on deep models.
- Flexibility and Customization: The unified representation allows for control of the level of complexity and processing based on system requirements.
- Practical Applicability: Suitable for environments that require a balance between processing speed and understanding accuracy, such as visual aids and automated classification systems.

VI. Conclusion and Future Work

A. Limitations

Although the proposed hybrid approach shows promising potential for improving caption quality, there are some aspects that require in-depth future study. These include the possibility that using two steps might make the process more complicated, as well as how much the quality of the first descriptions created with traditional templates affects the results. However, these limitations can be mitigated by improving the design of the initial templates or adopting modern techniques to control the overlap between the two stages, paving the way for a more effective and efficient development of the hybrid model.

B. Conclusion

In this paper, traditional and modern techniques for generating image captions are reviewed and analyzed. The paper demonstrates how modern techniques, particularly deep learning-based models, improve the quality of captions compared to traditional methods. CNNs were studied as encoders, along with transformer-based models and self-attention mechanisms, to improve the diversity and quality of captions. Results from the literature review indicate that the use of modern models contributes to a 15% to 30% improvement in performance indicators such as BLEU and CIDEr compared to traditional approaches, due to their ability to understand visual and semantic context more deeply.

One of the most important contributions highlighted by the paper is the theoretical introduction of a hybrid approach that combines traditional and modern methods, using a fusion strategy. Traditional methods are used to generate rapid and highly efficient initial descriptions, while modern models refine these descriptions and add deeper contextual details. Based on theoretical estimates based on previous studies, this approach is expected to improve computational efficiency by approximately 20% compared to using deep models alone, without compromising caption accuracy. It may even improve it by 10–15% thanks to the balanced contextual combination of the two methods.

Despite this progress, challenges remain, particularly regarding the quality of the captions and their consistency with human descriptions. This is due to the limited availability of datasets, which may not fully reflect the diversity and complexity of real-world scenes. This calls for improved data representation and the development of more effective contextual matching mechanisms in generative models.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Nuha .M. study concepts study design and analysis and wrote the research; Nada H. provided critical revisions contributed to the theoretical framework and reviewed the manuscript; all authors approved the final version.

References

[1] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–39, 2023.

[2] W. Hussain and A. Ghani, "Applying similarity measures to improve query expansion," *Iraqi J. Sci.*, pp. 2053–2063, 2021.

[3] Y. Ushiku, M. Yamaguchi, Y. Mukuta, and T. Harada, "Common subspace for model and similarity: Phrase learning for caption generation from images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2668–2676.

[4] L. Wu, "Generating descriptive and accurate image captions with neural networks," Ph.D. dissertation, University of Technology Sydney (Australia), 2019.

[5] A. Naseri and N. H. M. Ali, "Detection of drones with YOLOv4 deep learning algorithm," *Int. J. Nonlinear Anal. Appl.*, vol. 13, no. 2, pp. 2709–2722, 2022.

[6] Y. Wang, J. Xu, Y. Sun, and B. He, "Image captioning based on deep learning methods: A survey," arXiv preprint arXiv:1905.08110, 2019.

[7] P. Mathur, "A survey on various deep learning models for automatic image captioning," in *Proc. J. Phys. Conf. Ser.*, vol. 1950, no. 1, 2021, p. 012045.

[8] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013.

[9] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist.*, 2014, vol. 2, pp. 592–598.

[10] J. Devlin *et al.*, "Exploring nearest neighbor approaches for image captioning," arXiv preprint, arXiv:1505.04467, 2015.

[11] V. Ordonez *et al.*, "Large scale retrieval and generation of image descriptions," *Int. J. Comput. Vis.*, vol. 119, pp. 46–59, 2016.

[12] N. Krishnamoorthy *et al.*, "Generating natural-language video descriptions using text-mined knowledge," in *Proc. AAAI Conf. Artif. Intell.*, vol. 27, no. 1, 2013, pp. 541–547.

[13] G. Kulkarni *et al.*, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[14] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2013, pp. 1292–1302.

[15] R. Lebret, P. Pinheiro, and R. Collobert, "Phrase-based image captioning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2085–2094.

[16] R. Xu, C. Xiong, and W. Chen, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, 2015.

[17] Y. Ming *et al.*, "Visuals to text: A comprehensive review on automatic image captioning," *IEEE/CAA J. Autom. Sin.*, vol. 9, no. 8, pp. 1339–1365, Aug. 2022.

[18] Y. Ma *et al.*, "Towards local visual modeling for image captioning," *Pattern Recognit.*, vol. 138, p. 109420, Jun. 2023.

[19] S. S. Hussein, "Foreground object detection and separation based on region contrast," *Iraqi J. Sci.*, pp. 1963–1969, 2017.

[20] K. F. Mohammad, S. Sadiq, and M. S. Islam, "Improved bengali image captioning via deep convolutional neural network based encoder-decoder model," in *Proc. Int. Joint Conf. Adv. Comput. Intell.*, Springer, 2021, pp. 217–229.

[21] O. Shinde, R. Gawde, and A. Pardkar, "Image caption generation methodologies," *Int. Res. J. Eng. Technol.*, vol. 8, no. 4, pp. 2395–0056, 2021.

[22] X. Zhang *et al.*, "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15465–15474.

[23] S. Datta *et al.*, "Align2Ground: Weakly supervised phrase grounding guided by image-caption alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2601–2610.

[24] X. Chen *et al.*, "Leveraging unpaired out-of-domain data for image captioning," *Pattern Recognit. Lett.*, vol. 132, pp. 132–140, Apr. 2020.

[25] L. Yang *et al.*, "Object relation attention for image paragraph captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3136–3144.

[26] L. Guo *et al.*, "Aligning linguistic words and visual semantic units for image captioning," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 765–773.

[27] M. Stefanini *et al.*, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2022.

[28] T. Yao *et al.*, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 684–699.

[29] X. Yang, H. Zhang, and J. Cai, "Auto-encoding and distilling scene graphs for image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2313–2327, May 2020.

[30] W. Zhang *et al.*, "Consensus graph representation learning for better grounded image captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3394–3402.

[31] K. Nguyen *et al.*, "In defense of scene graphs for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1407–1416.

[32] N. Funckes. (2020). Tag: Automated image captioning. [Online]. Available: https://scholarworks.gvsu.edu/cgi/viewcontent.cgi?article=1005&context=mcnair_manuscripts

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

[34] Q. Wang and A. B. Chan, "CNN+CNN: Convolutional decoders for image captioning," arXiv preprint, arXiv:1805.09019, 2018.

[35] L. Wu *et al.*, "Recall what you see continually using grid LSTM in image captioning," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 808–818, Mar. 2019.

[36] Z. Song *et al.*, "Image captioning with context-aware auxiliary guidance," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2584–2592.

[37] R. Khan *et al.*, "A deep neural framework for image caption generation using GRU-based attention mechanism," arXiv preprint, arXiv:2203.01594, 2022.

[38] H. Zhou *et al.*, "Self-learning for few-shot remote sensing image captioning," *Remote Sens.*, vol. 14, no. 18, p. 4606, Sep. 2022.

[39] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[40] D. Cao *et al.*, "Cross-modal recipe retrieval via parallel-and cross-attention network learning," *Knowl.-Based Syst.*, vol. 193, 105428, Apr. 2020.

[41] F. Liu *et al.*, "Exploring and distilling cross-modal information for image captioning," arXiv preprint, arXiv:2002.12585, 2020.

[42] K. E. Zhang *et al.*, "Cross-modal image sentiment analysis via deep correlation of textual semantic," *Knowl.-Based Syst.*, vol. 216, 106803, Mar. 2021.

[43] M. Pourkeshavarz *et al.*, "Stacked cross-modal feature consolidation attention networks for image captioning," *Multimedia Tools Appl.*, pp. 1–25, 2023.

[44] S. Wu *et al.*, "Scene attention mechanism for remote sensing image caption generation," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.

[45] N. T. Tran, "Enhancing accuracy for classification using the CNN model and hyperparameter optimization algorithm," *Indones. J. Electr. Eng. Inform.*, vol. 12, no. 3, 2024.

[46] Y. Pan *et al.*, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10971–10980.

[47] X. Zhu *et al.*, "Captioning transformer with stacked attention modules," *Appl. Sci.*, vol. 8, no. 5, p. 739, 2018.

[48] J. Yu *et al.*, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, Dec. 2019.

[49] S. Herdade *et al.*, "Image captioning: Transforming objects into words," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[50] L. Guo *et al.*, "Normalized and geometry-aware self-attention network for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10327–10336.

[51] W. Liu *et al.*, "CPTR: Full transformer network for image captioning," arXiv preprint, arXiv:2101.10804, 2021.

[52] C. Sundaramoorthy *et al.*, "End-to-end attention-based image captioning," arXiv preprint, arXiv:2104.14721, 2021.

[53] K. I. *et al.*, "Image captioning using motion-CNN with object detection," *Sensors*, vol. 21, no. 4, p. 1270, 2021.

[54] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[55] Z. Zhang *et al.*, "Psalm: Pixelwise segmentation with large multi-modal model," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2024, pp. 74–91.

[56] J. Qiu *et al.*, "Visual-based policy learning with latent language encoding," in *Proc. ICML 2023*, 2023.

[57] X. Li *et al.*, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, 2020.

[58] H. Maru *et al.*, "Comparison of image encoder architectures for image captioning," in *Proc. Int. Conf. Comput. Methodol. Commun.*, 2021, pp. 740–744.

[59] N. Zeng *et al.*, "Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip," *Neurocomputing*, vol. 425, pp. 173–180, Jan. 2021.

[60] A. Liu *et al.*, "Multi-level policy and reward reinforcement learning for image captioning," in *Proc. IJCAI*, 2018, pp. 821–827.

[61] P. H. Seo *et al.*, "Reinforcing an image caption generator using off-line human feedback," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 3, 2020, pp. 2693–2700.

[62] U. Honda *et al.*, "Switching to discriminative image captioning by relieving a bottleneck of reinforcement learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1124–1134.

[63] W. Wang *et al.*, "ViLTA: Enhancing vision-language pre-training through textual augmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3158–3169.

[64] L. Zhou *et al.*, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13041–13049.

[65] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP prefix for image captioning," arXiv preprint, arXiv:2111.09734, 2021.

[66] J. Li *et al.*, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 12888–12900.

[67] J.-B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 23716–23736, 2022.

[68] W. Jin *et al.*, "GRILL: Grounded vision-language pre-training via aligning text and image regions," arXiv preprint, arXiv:2305.14676, 2023.

[69] L. Chen *et al.*, "ShareGPT4V: Improving large multi-modal models with better captions," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2024, pp. 370–387.

[70] J. Gao *et al.*, “Self-critical n-step training for image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6300–6308.

[71] Z. Ren *et al.*, “A mask-guided transformer network with topic token for remote sensing image captioning,” *Remote Sens.*, vol. 14, no. 12, p. 2939, 2022.

[72] F. Daneshfar, A. Bartani, and P. Lotfi, “Image captioning by diffusion models: A survey,” *Eng. Appl. Artif. Intell.*, vol. 138, p. 109288, 2024.

[73] C.-W. Kuo and Z. Kira, “Haav: Hierarchical aggregation of augmented views for image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11039–11049.

[74] Z. Zeng *et al.*, “Conzic: Controllable zero-shot image captioning by sampling-based polishing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23465–23476.

[75] Z. Zeng *et al.*, “MEACAP: Memory-augmented zero-shot image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 14100–14110.

[76] P. Yamini, F. Daneshfar, and A. Ghorbani, “KurdSM: Transformer-based model for Kurdish abstractive text summarization with an annotated corpus,” *Iran. J. Electr. Electron. Eng.*, vol. 20, no. 4, 2024.

[77] F. Daneshfar *et al.*, “Elastic deep multi-view autoencoder with diversity embedding,” *Inf. Sci.*, vol. 689, 121482, 2025.