Deep Image Captioning Survey: A Resource Availability Perspective

Mousa Al Sulaimi
Kuwait University
Kuwait
Mausa.alsaleemi@grad.ku.edu.kw

Imtiaz Ahmad, Mohammad Jeragh
Kuwait University
Kuwait
{imtiaz.ahmad , mohammad.jeragh}@ku.edu.kw

Abstract—Recent advances in deep learning have enabled machines to see, hear, and even speak. In some cases, with the help of deep learning, machines have also outperformed humans in these complex tasks. Such improvements have reignited interest in many fields. Image captioning, which is considered an intersection between computer vision and natural language processing, has recently received significant attention. Deep learning-based image captioning models represent a great improvement on traditional methods. However, most of the work done in image captioning is based on supervised deep learning methods. Recently, unsupervised image captioning has started to gather momentum. This paper presents the first survey that focuses on unsupervised and semi-supervised image captioning techniques and methods. Additionally, the survey shows how such methods can be used with different data availability and data pairing settings, where some methods can be used with paired data, while others can be used with unpaired data. Furthermore, special cases of unpaired data, such as cross-domain and cross-lingual image captioning, are also discussed. Finally, the survey presents a discussion on challenges and future research directions of image captioning.

I. Introduction

Images play an import role in our lives. They are found in homes, streets, workplaces ,and online. They are used for instructions, like stop signs in the street, or an art piece hung in behind a protective glass in the Louvre museum, or maybe for entertainment, such as the edited images found online. Regardless of their purpose, all images have a common role: communication.

The internet has made information accessible to everyone. Data, and thus information, flows to a person through a search engine within seconds of clicking the enter button. However, a large amount of information is not yet fully accessible, comparable, or searchable with the use of text inquiries as keywords. For example, textual information can be easily compared with requested information and thus retrieved. Unfortunately, the same cannot be said for images and videos [1]. This is because computers cannot convert an image or a video into a textual description to compare text received in an inquiry without preprocessing.

Image captioning is the task of creating textual descriptions of images [2]. One of the earliest and simplest methods used in image captioning is retrieval-based image captioning [2]. With this method, the system has a database of images and their captions or discerptions. When an inquiry is received by the system, it checks for an exact or close image match. The description of the matched image is then retrieved. Since the key criterium in this method is image similarity, research at that time focused on different measures of image similarity to find the best match. However, since the descriptions are directly retrieved from a predefined database, sentences produced using retrieval based methods are not diverse and rarely return the correct image description [2]. Another common method of early image captioning was template-based captioning [2], whereby the system analyzes an image for visual concepts and then links those concepts with a predefined sentence template in the system. Although this method generates a more relevant description than retrieval-based methods [2], it still generates a rigid selection of sentences for images with the same visual concepts, with little diversity.

Retrieval-based and template-based image captioning methods were the go-to methods for a long time. However, this changed with the appearance of deep learning and neural networks, especially after deep learning enhanced computer vision with the use of convolutional neural networks (CNN) [3] and enabled neural machine translations (NMT) [4] using long-short-term-memory (LSTM). Since image captioning is situated at the intersection of computer vision and natural language processing [5], it made sense to combine CNN and LSTMs to generate image captions. Indeed, published research shows that this new approach outperforms retrieval-based and template-based methods [5]. Further significant improvement was achieved when attention [6] mechanisms began to be used with image captioning in 2015, thus attracting more research on image captions.

Despite these great improvements in image captioning, most research on the subject has been focused on supervised image captioning. Only recently has the widespread use of generative adversarial networks (GANs) [7] helped to pave the way for research on unsupervised image captioning [2]. Recently, such GAN based models were proposed to perform end-to-end unsupervised image captioning [8], [9] and improved image captioning [10], [11] in an unsupervised manner. Additionally, some researchers have proposed using semi-supervised techniques to relax the restriction of fully labeled data. The surveys [2], [12-15] group and present supervised methods used for image captioning, alongside the various configurations and techniques used for this purpose. Meanwhile, [2] have touched on the subject of unsupervised machine learning. Nevertheless , none of these surveys have discussed unsupervised or semi-supervised image captioning in detail. Since research on unsupervised and semi-supervised

image captioning has started to gain momentum, a survey analyzing and presenting the techniques of unsupervised and semi-supervised image captioning is required to enable subsequent research to build on the existing work.

Based on the above discussion, this paper presents a survey of image captioning using unsupervised and semi-supervised techniques. To the authors' best knowledge, this is the first attempt to provide a comprehensive survey of unsupervised and semi-supervised image captioning techniques. Since data availability is usually a key factor in deciding which training techniques to use, this paper integrates the data availability perspective into the choice of unsupervised and semi-supervised techniques in the context of image captioning, in an attempt to highlight the unsupervised and semi-supervised image captioning techniques used in various data availability scenarios.

The main contributions of the current paper are to:

1) Offer an overview of recent advances in image captioning in general.
2) Provide a detailed analysis unsupervised technique in image captioning.
3) Provide a detailed analysis of methods for paired and unpaired data.
4) Present future work and open challenges.

The rest of this paper is organized as follows. Section II presents related work, discusses image surveys, and compares them with the survey presented in this paper. Section III describes image caption taxonomy and comparison method used in this survey. Section IV discusses and distinguishes unsupervised image captioning techniques. Section V explores semi-supervised image captioning. Section VI summarizes open challenges and issues, Section VII highlights the limitation in the current survey and finally, section VIII concludes this paper.

II. Related work

Interest in the field of image captioning has grown at a rapid pace in the past few years. The current understanding of image captioning can be classified according to various perspectives. This section presents previously published surveys. Their shortcomings are discussed, and they are compared with this proposed survey.

The work done in [12] categorizes image captioning according to two categories. The first category includes traditional methods such as retrieval-based and template-based methods, while the second category includes deep learning methods. Additionally, the authors examine different types of image captioning improvements, such as attention mechanisms and different encoder/decoder configurations. The authors also discuss the available datasets and the evaluation metrics that are currently used to evaluate the image captioning methods. In [13] the authors survey image captioning, classify its approaches into three different categories, and discuss their advantages and disadvantages. The first and second categories are retrieval-based and template-based methods, respectively, while the third category is neural network-based, for which the authors discuss different models and architectures. Additionally, the authors compare several state-of-the-art methods and present the results based on multiple benchmarks. Finally, the authors discuss challenges and open issues in the field of image captioning. It is worth noting that the authors of [13] discuss a number of semi-supervised models. However, the paper does not provide a comprehensive study of semi-supervised and unsupervised image captioning. Similarly, [14] follows the same classification approach as [12], [13] but discusses supervised techniques only. The survey conducted in [15] classifies image captioning models with either a generation problem or a retrieval problem. The authors also discuss various datasets and evaluation metrics that are currently used for image captioning. In the challenges section, the authors propse unsupervised image captioning as a subject for future work and a source of open problems. A recent survey [2] provides a detailed analysis of machine learning-based image captioning techniques. Similar to the other surveys, this study divides image captioning into traditional machine learning-based and deep learning-based methods. Additionally, this survey touches on the subjects of unsupervised and semi-supervised image captioning. Finally, similar to [15], the authors of [2] indicate the need to perform image captioning on unlabeled datasets as an open issue.

As shown in Table I, most existing surveys on image captioning have focused on supervised learning methods, while semi-supervised and unsupervised methods have received less attention. Additionally, none of the previously published studies have discussed image captioning from the data availability or data pairing perspectives. Data availability and paring are key factors for deciding which techniques to use. A survey analyzing and presenting the techniques of unsupervised and semi-supervised image captioning in the context of data availability and pairing is required to enable subsequent research to build on existing work.

III. Image captioning taxonomy and comparison method

Image captioning refers to the task of generating a natural sentence that reflects the visual content of an image. This is associated with accurately expressing image content in a sentence, which is known as "translation" from image domain to the language domain . The main challenges in providing an accurate description lie in recognizing objects, attributes, and activities in an image, in addition to the establishment of fluent sentences which satisfy grammatical constraints and rules that are natural, diverse, and indistinguishable from human captioned sentences.

There are many methods that can be used to perform image captioning. These methods can be classified and compared from different aspects. For example, the work in [2] compares methods base on architecture and topology. Technique type is another aspect that is used to compare existing published work. Similarly, method supervision, such as considering if a method is supervised or not. Is a key factor used in the method selection. Especially when the needed data may not be fully labelled in the desired language or domain. Another closely related factor is data parity. Like supervision, comparing methods based on data parity is also important when data is scarce. Especially that image captioning requires two types of data modalities. The first is the images, and the second is the language or sentences. And both datasets can be in different

pairing conditions. For example, if the image dataset is related to the sentence dataset then both datasets are said to be paired. Conversely, if the image dataset is not related to the sentence dataset then both datasets are said to be unpaired. However, there are some cases where the sentence dataset and image datasets may have an indirect relationship. In this case the datasets are referred to as semi-paired datasets.

The survey presented in this paper classifies and compares existing published work on the data availability and parity in addition to the type of supervision used in each technique with a focus on unsupervised methods. Fig. 1 presents a hierarchy of the image captioning techniques classification adopted in the current survey. It is worth noting that supervised image captioning techniques are not covered in current survey. Work in [2,12-15] provide a detailed analysis and comparison of supervised and traditional image captioning techniques.

TABLE I. COMPARISION BETWEEN DIFFERENT SURVEYS

	Learning Methods							Data Availability		
Surveys	Traditional Methods	Deep Learning			Improvement Techniques			Paired Data	Semi-Paired Data	Unpaired Data
		Supervised	Semi-Supervised	Unsupervised	Supervised	Semi-Supervised	Unsupervised			
[12]	√	√			√			√		
[13]	√	√	√		√			√		
[14]	√	√			√			√		
[15]	√				√			√		
[2]	√	√	√	√	√			√		√
Current Survey			√	√		√	√	√	√	√

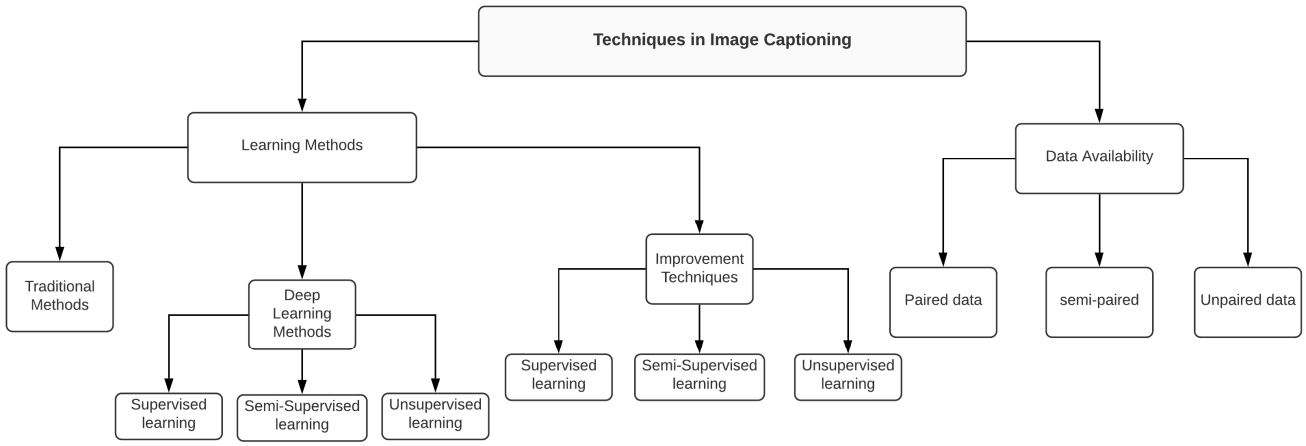

Fig 1. Classification of Image caption techniques

IV. UNSUPERVISED TEHCNIQUES IN IMAGE CAPTIONING

Unsupervised techniques in image captioning began to receive research attention in recent years. These techniques are commonly used in cases without image–sentence pairs, when attempting to improve an image captioning model beyond the capabilities of the available data, or when the available data is out of the image captioning domain. The following sections

describe the techniques used in existing work to address each of the aforementioned cases.

A. Unsupervised image captioning

Current state-of-the-art image captioning techniques require a paired image–sentence corpus. Where each image has at least one ground truth caption. This requires a large amount of tedious manual work to create. Such work might not be feasible. As a result of this limitation, image captioning research has focused on the English language.

This problem has recently started gaining researchers' attention. And various methods were proposed to perform image captioning without the need to have labelled data. The challenge stems from the fact that the image captioning model should be capable of ensuring that visual concepts found in images can be aligned with words in sentences generated by the model, without having access to the ground truth during training. Therefore, maximum likelihood loss cannot be applied directly without prior processing. Additionally, it is worth noting that images and sentences (text) are of different modalities and different data types. For example, words are discrete, whereas images are continuous. Hence, models need to be able to cater for both data types without losing important information.

Yang et al. [8] attempt to address the problem of image captioning without the need for a paired image–sentence corpus. In order to perform this unsupervised image captioning, the authors propose a new model based on GAN in which an encoder–decoder captioner is used to generate captions. A discriminator is responsible for checking that captions generated by the model are syntactically correct, describes the visual concepts in the image, and verify that all visual concepts are captured by the caption. Based on these three criteria, the model is either rewarded or penalized. Additionally, the authors coupled the captioner and discriminator in a way that allows the latter to reconstruct the image features to ensure better alignment. In the model proposed by [8], pre-training of the generator and discriminator is required for the model to return good results. This model is one of the first attempts to address the problem of unsupervised image caption. It is noted that in this model the discriminator is responsible for governing the training of the entire model. Which requires backpropagation between the discriminator and generator.

Backpropagation is problematic in the case of non-differentiable discrete data. Researchers often use an approximation technique such as the Gumbel softmax approximation and reinforcement trick/rule to perform backpropagation on discrete data. The authors of [9] attempt to avoid this issue by proposing a different model that comprises two components. The first component is a language model that is trained on unpaired sentences to learn a structured visual concept word embedding. In this language model, sentences with overlapping visual concepts are embedded close to each other. This is achieved by adopting a special embedding method, as described in [9]. The second component is a fully connected neural network that acts as a visual alignment component. It aligns visual concepts extracted from images and projects them to the same structured embedding learnt by the language model. To ensure that the visual concepts are correctly captured, the authors propose a discriminator network to compare the embedding learnt with the visual elements extracted from the training images. It is essential to note that, unlike [8], the discriminator in this model considers continuous differentiable feature space, thus, backpropagation is feasible. The discriminator decision is fed back using policy gradient updates to the alignment model. It is also worth noting that the decoder of the language model generates the output sentences by decoding the learnt embedding and the learnt alignment. In other words, unlike generators used in GANs' architecture, the decoder of the language model does not see the actual visual concepts to generate a correct sentence.

The work done in [33], attempt to further simplify unsupervised image captioning models. Where the authors propose an attention based recurrent relational memory to align images and sentences without using adversarial training or reinforcement learning. The proposed model is composed of three memory-based networks each taking a separate role. Where the first network is a memory encoder, the second is a memory decoder, and the third as an embedding reconstructer.

TABLE II. COPARIOSN BETWEEN MODELS PROPOSED BY [8] AND [9] UNDER THE MSCOCO [34] DATASET

Method	Objective	Requires same domain data	Requires pseudo pairs	BLEU 4	METEOR	ROUGE	CIDEr	SPICE
GANs with multi rewards [8]	Unsupervised image captioning	No	Yes	5.6	12.4	28.7	28.6	8.1
shared multimodal networks [9]	Unsupervised image captioning	No	Yes	6.5	12.9	35.1	22.7	7.4
Recurrent relational memory [33]	Unsupervised image captioning	No	Yes	8.3	14.0	35.0	29.3	9.6

As shown in Table II, the model proposed by [9] shows improved results compared to the model proposed by [8]. This can be attributed to the dedicated domain alignment mechanism introduced by [9], which was trained using adversarial training with the use of a discriminator for alignment explicitly on continuous data. In contrast, [8] used adversarial training to train the entire model. Such training would require performing backpropagation on discrete data and back through an LSTM before adjusting the image alignment. The same observation can be made when comparing [33] with [8,9]. It is obvious that the sophisticated attention mechanism used in [33] to aligned image and sentence data helped improve the model's performance on all metrics.

Desipite the fact that unsupervised image captioning is still in its early stages, its proposed models are growing in complexity as noted by [33]. Additioanly, when compared with state-of-the-art supervised image captioning models, unsupervised image captioning exhibits significantly poorer performance on available metrics. Therefore, models accuracy and complexity are open issues that needs to be addressed in this area .

B. Unpaired image captioning

Unpaired image captioning is like unsupervised image captioning with a one small exception. In unsupervised image captioning both the image dataset and language training dataset are assumed to be from different domains. Where in unpaired image captioning the image dataset and the language are assumed to be from the same domain. That is both the images and sentences that the model trains on describes similar or the same concepts and objects. This simplifies the unsupervised image captioning problem. It is worth noting, that although both images and sentence are from the same domain. They are not grounded together as pair.

The first to propose unpaired image captioning was the work done in [25]. Where the authors proposed a model to perform English language image captions using an image captioning model trained on the Chinese language. In the proposed model the captioner first generates captions in Chinese. The captions are then passed to a neural machine translation (NMT) model that is trained on Chinese-to-English translation. The goal of the NMT model is to translate the Chinese captions to English. To ensure good translation, the authors use ta sentence auto-encoder that is trained on syntactically correct sentences English sentences.

The work done in [25] assumes the existence of a paired image-sentence dataset in a one language to perform unpaired captioning in another. Such assumption is not always true. To eliminate this assumption, the work done in [31] follows a different approach for unpaired image captioning. In [31] the authors propose a model based on graph align method. Where two graph generators are used to generate scene graphs for both images and unpaired sentences. The graphs are then encoded in a graph autoencoder network [32] to be used in the captioning model. To perform the image captioning, a language model is trained on sentence scene graph to generate captions. During inference, the image scene graph is passed to the language model to generate captions. Additionally, the authors propose using GANs to align the image and sentence scene graphs before inference to produce valid captions.

Additional to the above-described methods, it is worth mentioning that since unsupervised image captioning models make less assumption about the data. Such models can also be used in unpaired image captioning. Table III shows a comparison between models proposed for unpaired image captioning.

TABLE III. COMPARISON OF THE METHODS PRPOSED BY [8, 9, 25, 31]

Method	Objective	Require data in domain	Requires pseudo pairs	BLEU 4	METEOR	ROUGE	CIDEr	SPICE
GANs with multi rewards[8]	Unpaired image captioning	No	Yes	18.6	17.9	43.1	54.9	11.1
shared multimodal networks [9]	Unpaired image captioning	No	Yes	19.3	20.1	45.4	63.6	12.8
Pivot language [25]	Unpaired image captioning	Yes	Yes	5.4	13.2	-	17.7	-
Graph Align method [31]	Unpaired image captioning	Yes	Yes	21.5	20.9	47.2	69.5	15.0

C. Image captioning improvements without additional data

Supervised image captioning exhibits high performance when compared with other techniques. In fact, all state-of-the-art models are supervised. However, it is noticeable that image captions generated from a supervised image captioning seems artificial and less natural. This observation is attributed to the limited vocabulary used in image caption corpora and the maximum likelihood techniques used in supervised image captioning, which restrict the generated samples to be similar to the training data [10]. Therefore, to create expressive, diverse, and natural image captioning with the use of supervised techniques, it is essential to create larger datasets with more diverse captions for each image. Alternatively, unsupervised image captioning techniques do not require labeled datasets. Accordingly, such techniques can be used to add naturalness and diversity to image captioning without additional data labels.

The research conducted in [10] attempts to address the issues of image caption diversity and naturalness. Where [10] proposes a model that is capable of generating diverse image captions and human-like captions that cannot be differentiated from human-crafted captions. Unlike existing models, the authors in [10] propose a single model that generates multiple correct captions describing the same image. This is done by employing GANs. In the proposed model, the generator comprises an encoder–decoder architecture that produces multiple captions for the same image. Meanwhile, the discriminator examines the generated captions and the captioned image to decide whether the captions are real and relevant to the image. It also checks whether the multiple captions generated for the same image are diverse. Subsequently, the discriminator's decision is fed back to the generator. It is worth noting that the authors of [10] have recognized that adversarial models do not perform well when evaluated under automatic correlation metrics such as BLEU, SPICE, and METEOR, because such evaluation techniques prefer frequently used n-grams, thus discouraging diversity. However, the experiment in [10] showed that human inspection favors adversarial captions.

While paper [10] focuses on generating image captions that are correct and diverse, the authors of [11] extend the problem by adding a third objective, namely naturalness. The authors propose a GAN-based model similar to the one used in [10]. However, while the model in [10] uses multiple captions generated for the same image to increase the diversity, the discriminator in [11] uses captions generated across the entire training mini-batch in order to learn image–caption relevancy. The authors extended their proposed model to create image paragraph generation by utilizing hierarchical LSTMs in the generator and discriminator. Similar to [10], the authors of [11] discuss the inadequacy of existing evaluation metrics. And introduce two new metrics that are more suitable for GANs, namely E-GAN and E-NGAN.

While GANs are popular unsupervised deep learning models, they are generally hyperparameters-sensitive. Thus, training them requires some trial and error. Several solutions to this problem have been proposed. The most prominent is a specific type of GANs called a Wasserstein-GAN [36] or WGAN, such as that used in [9]. The authors of [17] proposes a model that produces diverse captions and does not suffer from the training instability found in GANs. The proposed model is similar to GAN-based models of [10], [11]. However, in a typical GAN setup the generator and discriminator take turns in teaching each other. This is not the case in the model proposed by [17], in which the discriminator is pretrained and takes part in training the generator, while the latter does not impact the discriminator, as it remains frozen after pretraining. In the proposed model, the discriminator is a neural image retrieval system pretrained on the ground truth, it receives an image and its caption from the generator model and returns loss and gradient updates to the generator. Thus, the generator is trained based on the feedback of the discriminator.

The work presented in [18] follows a different approach to enhance the results and performance of discrete GANs. The paper proposes sophisticated generators and discriminators equipped with attention mechanisms. The generator is extended with an adaptive attention mechanism, which includes visual components from the training image and text components from the training sentence. The discriminator is also equipped with an attention mechanism used to assess the similarity between an image and its generated caption. This is done by using the image feature extracted directly out of the CNN's pooling layer in the generator and word embeddings generated by the language model. To ensure the stability of GAN training, the discriminator returns its feedback to the generator to train using the self-critical sequence training method [19]. This method is known to stabilize and normalize generator training, as well as solving the problem of non-differentiable gradients.

Table IV compares the above-mentioned methods in terms of several metrics on the MSCOCO dataset. It important to reemphasize the conclusion made by [18] and [11], which states that current automatic evaluation metrics do not favor diversity. Therefore, a new diversity measure must be created and adopted to be able to fairly compare diversity models.

D. Cross-domain image captioning

The previously discussed work focuses on improving the diversity, naturalness, and accuracy of image captioning models in an unsupervised fashion. Meanwhile, other researchers have focused on other problems and addressed them by employing unsupervised techniques. One of these new and unexplored problems is the issue of domain shift and cross-domain captioning. This problem arose due to the observation that image captioning datasets are biased towards certain domains. For example, the MSCOCO dataset is human centric and most of its images relate to humans and their activities. In contrast, the Oxford-102 dataset [35] concerns shapes and the colors of flowers. Generating a balanced, unbiased dataset would require significant amounts of time and resources. This leads to the attractive notion of building models that can learn captioning on one domain and transfer what is learned to another domain.

TABLE IV. COMPARISON OF THE METHODS PRPOSED BY [10, 11, 17, 18]

Method	Dataset (training/ validation)	Requires separate diversity data	BLEU 4	METEOR	ROUGE	CIDEr	SPICE
Adversarial Training [10]	MSCOCO	No	-	27.2	-	-	18.7
E-GAN [11]	MSCOCO	No	20.7	22.4	47.5	79.5	19.2
Adversarial Training with pertained discriminator [17]	MSCOCO	No	32.0	26.2	54.5	103	19.7
Adversarial Semantic Alignment [18]	MSCOCO	No	-	27.1	-	111.1	-

The work presented in [21] explores the idea of unsupervised cross-domain image captioning. It introduces a model trained on one domain that can transfer its knowledge to another domain. Like most of the models discussed in this section, the proposed model is based on GAN with a standard generator. The discriminator comprises two subnetworks. The first network is called the domain critic. It is responsible for assessing any caption generated by the generator, and classifying whether it falls within the source domain, target domain, or generated. The second subnetwork in the discriminator is the multi-modal critic. The responsibility of this critic is to receive the image caption generated by the generator along with the captioned image, and assess the relevancy of the image as paired, unpaired, or generated. The goal of the generator is to generate paired and target captions. In the proposed model, the generator and multi-modal critic are pretrained on source domain data, while the domain critic is trained on target domain sentences.

Similar to [21], the work presented in [22], [23] addresses the problem of cross-domain image captioning. The authors of [22], [23] employ the dual learning technique [24]. However, instead of using two critic networks like [21], the proposed model is trained on two tasks; the first task is image captioning, and the second task is text–image synthesis. This is achieved by each having two subnetworks each dedicated for a single task. The subnetwork responsible for image captioning follows and uses a standard generator that is trained on the source domain image captions. The subnetwork used for text–image synthesis is a GAN network with a discriminator that checks if the image synthesized is correct or not. Additionally, the authors of [22], [23] propose that the two networks are further trained using a dual training game, whereby the image caption sees an image and generates a corresponding caption. That caption is fed to the text–image synthesis subnetwork to generate an image. The synthesized image is then compared with the original image by a discriminator to update both subnetworks. Finally, the authors use unpaired target domain images and sentences to fine-tune the proposed model to perform cross-domain image captioning.

Table V compares described techniques on different metrics and shows that using specialized discriminators, as done in [21], can produce good results. However, since the specialized discriminators in [21] are isolated from each other, it is possible that the impact of one discriminator may dominate the other, thus balancing the effect of competing discriminators. Using multi-task learning, as in [22], [23], eliminates the effect of competing discriminators, as each task networks feeds the other network with training data. Therefore, the accuracy of multi-task learning is higher than that of dual critic networks. With that said, it is important to note that learning to reconstruct and synthesize images in multi-task learning is computationally expensive.

E. Cross-lingual image captioning

As indicated in the previous sections, Due to the lack of resources, most of the image captioning research is done on the English language. Therefor it is difficult to use supervised image captioning to generate captions in different languages. Alternatively, and like cross-domain image captioning unsupervised image captioning techniques can be used to transfer some of the information found in English datasets to perform image captioning in other data.

The work done by [25] is considered to be one of the first models that performs image captioning without having a paired dataset in the target language. The authors proposed a model that generates English image captions with the use of an existing Chinese language dataset. The model constitutes of an encoder–decoder Chinese language captioner that is trained on a Chinese language image captioning dataset. The generated captions are then passed to a neural machine translation (NMT) model that is trained on Chinese-to-English translation. The goal of the NMT model is to translate the Chinese captions into English. However, the authors of [25] claim that performing direct translation of captions introduces two issues: first, translation errors are propagated to captions; and second, it is noticed that target and source languages are from different domains and distributions. Thus, the authors propose that the

translated English captions are curated by a sentence auto-encoder that is trained on syntactically correct sentences. The auto-encoder receives the translated English image captions and correct any errors in the sentence. It is worth mentioning that the model described by [25] requires two types of datasets; the first type is an image–sentence pair that is used for the image captioning model. The second comprises source–target language pairs to perform NMT.

TABLE V. COMPARISON OF DIFFERENT TECHNIQUES USED IN CROSS-DOMAIN IMAGE CAPTIONING

Method	Dataset (source domain/ validation)	Requires separate target domain sentences	BLEU 4	METEOR	ROUGE	CIDEr
Dual learning [21]	MS-COCO	Yes Oxford-102	60.5	36.4	72.1	29.3
Multi-task Learning [22]	MS-COCO	Yes Oxford-102	71.6	43.0	82.4	79.7
Vibrational Auto-encoder with Additive Gaussian prior [23]	MS-COCO	Yes Oxford-102	73.5	46.1	84.5	90.6

Other researchers place greater emphasis on curating the translated image captions dataset. For example, in [26] the author proposes to translate the image captioning dataset before training the model. Therefore, a Chinese image captioning dataset is first translated into English. A supervised state-of-the-art encoder–decoder image captioning generator is then trained on the translated dataset with the maximum likelihood technique (MLE). However, since the captions have been automatically translated, it is assumed that some captions may be irrelevant to the image or syntactically incorrect. To address this, the generated captions are passed to a multimodal discriminator network, which receives the generated captions, along with the captioned image, and checks the relevancy of the generated captions with the input image by computing multi-level relevancy rewards. Additionally, it checks whether the generated caption seems natural by computing a fluency reward. Both rewards are then combined and fed back to the generator in order to update it for additional training.

While [25], [26] attempt to build models that curate imperfect datasets by introducing components that assess Caption–image relevancy and sentence fluency, the authors of [27] follow the opposite approach. The proposed model attempts to discard non-fluent training examples and ignores them during training. like [25], [26], the model proposed by [27] uses an encoder–decoder image caption generator which is trained on a translated image-captioning dataset. However, before passing a batch to train the model, the entire batch is assessed using a fluency guiding classifier. The classifier is neural network trained on fluent image captions using only the captions before and after the translations. Hence, when a training batch is prepared it is sent to the fluency classifier, which samples the batch for fluent captions, thus resulting in a high-quality training batch. The new training batch is then used to train an image captioning model. According to the work done in [27], human evaluation reveals that the captions generated by the model are preferred over those that are generated by the training model on the entire dataset. Table VI compares the three methods discussed in this section.

TABLE VI. COMPARISON OF DIFFERENT TECHNIQUES USED IN CROSS-LINGUAL IMAGE CAPTIONING

Approach	Training dataset	Validation dataset	BLEU 4	METEOR	CIDEr
Pivot language [25]	AIC-ICC	MSCOCO	5.4	13.2	17.7
Self-Supervised Rewards [26]	AIC-ICC	MSCOCO	11.1	14.2	79.5
Fluency-guided [27]	Flickr8k	Flickr8k-cn	24.1	-	47.6

V. SEMI-SUPERVISED IMAGE CAPTIONING

The previous sections discussed using different deep learning models and techniques to utilize labelled data and unlabeled data in unsupervised algorithms. Nevertheless, the previous sections did not discuss the case of partially labelled data. Discarding the partial labels will be a waste of data. Especially for data hungry models. Semi-supervised image captioning models can help in utilizing the partial labels to perform image captioning.

The authors of [28] attempt to perform semi-supervised image captioning or image captioning with partially labeled

data. The work in [28] proposes a method that utilizes a small portion of the labeled data to generate possible caption pairs for other unlabeled images. This is done by training a standard caption generator using a self-retrieval module. Which is a neural network trained on receiving a caption and retrieving a matching image from a dataset. During training, a generator captions an image and sends both the image and its caption to the self-retrieval network. The network then retrieves the image that best matches the caption from images of the current training mini-batch. The retrieval network calculates the difference between the similarity of the captioned image and the generated caption, and the similarity of the retrieved image and the generated caption. The difference between the two similarities is then backpropagated to the generator to be updated. Therefore, the generator is always trained to produce matching image–caption pairs. This method handles unlabeled data naturally as no ground truth data is required. However, since the retrieval network is pretrained and remains frozen during training of the generator, the generator quickly catches up with the retrieval network.

The work proposed by [29] is similar to [28], with the main difference that the retrieval network is replaced by a discriminator network which is trained on labeled image–caption pairs. When presented with an unpaired or generated image caption, the discriminator searches the space of captions and retrieves the best matching caption to compute the loss between it and the unpaired caption. The loss is then used to update the generator. In other words, the discriminator forms a new pseudo pair and treats it as the ground truth for training the model. However, it is important to note that the newly generated pairs are not noise free. The authors of [29] attempt to mitigate this by introducing a confidence score, which is used when calculating the loss. It is also worth noting that the generator and discriminator in this setup, unlike in [28], is in constant competition. This allows both networks to improve over time, thus resulting in a better model.

The models presented in [28], [29] are designed to utilize a small portion of fully labeled data to perform image captioning. While such a technique produces impressive results with labeled data as limited as 1% of the entire dataset, as shown in [28], they do not cater for unseen or novel objects, especially since both models in [28], [29] use candidate selection methods trained on labeled data. The work presented in [30] attempts to address this issue. The authors assume that single image labels are easier to acquire than image captions. Additionally, the authors of [30] argue that a single image label is nothing but an incomplete or partially labeled image caption. Based on this argument a novel training algorithm called partially specified sequence supervision (PS3) is proposed. The algorithm encodes the partially labelled sequence with the use of finite state automaton (FSA) and attempts to complete the missing part of the partially labeled sequence. The completed sequence, along with the rest of the labeled data, is then used to train the image captioning model. Therefore, this training method allows the image captioning model to be trained on datasets that contain image labels only, such as object detection datasets, which are known to have more objects than exist in traditional image captioning datasets. Table VII compares all three techniques.

TABLE VII. COMPARISON BETWEEN SEVERAL SEMI-SUPERVISED IMAGE CAPTIONING MODELS

Approach	Training dataset	Requires fully labeled data	Novel object captioning	Objective	METEOR	CIDEr
Self-Retrieval [28]	MSCOCO	No	No	Image captioning with partially labeled data	27.4	117.1
Adversarial Semi-Supervised learning [29]	MSCOCO	No	No	Image captioning with partially labeled data	29.4	125.5
Partially specified sequence using FSA[30]	MSCOCO	No	Yes	Novel object captioning	25.4	101.1

VI. Open issues and challenges

Great progress has been made in image captioning over the past few years. Captions generated from deep learning models are comparable to human-crafted captions. Semi-supervised and unsupervised deep learning techniques have

also allowed researchers to decrease their dependency on labelled datasets. In other cases, semi-supervised and unsupervised techniques have enabled researchers to relax the conditions of data pairing in a way that enables the use of different types of data from different domains.

Nevertheless, there is still room for improvement in the areas of image captioning, and particularly when using unsupervised techniques in image captioning. One less obvious issue encountered by researchers is the inadequacy of common performance measurement metrics such as BLEU, METEOR, SPICE, and CIDEr to reflect the performance of models, especially when diverse and natural captions are preferred. From [10], [18] it is evident that a sort of consensus has been reached among researchers, and that a new measure needs to be introduced to adequately measure the model's performance. Hence, this avenue is worth pursuing in the future to enable new models to be measured accurately.

Unsupervised and unpaired image captioning is still a challenge. This area has just recently attracted attention, but the results show that there is a noticeable gap between state-of-the-

art supervised models and unsupervised ones. The main challenge stems from the difficulty of aligning data of different modalities, such as text and images without a base or aligned examples. Some researchers have attempted to solve this issue by introducing pseudo pairs or loosely aligned data, such as [8, 9], while others have used no alignment at all, such as [32]. In addition to model's performance, the complexity of such unsupervised models is growing [33]. It would be interesting to see how other methods contribute in the future work.

Additionally, it is noted that caption naturalness is an issue in both supervised and unsupervised captioning. One aspect of caption naturalness that has recently been central in supervised image captioning research is the non-factual description of images, sometimes referred to as image caption stylization. In this case, image captions not only need to describe all factual components of an image or a scene, but also add some personality to the caption, such as humor or optimism. To address this issue in supervised image captioning, additional data needs to be collected with stylized data. One alternative to collecting new data would be the use of unsupervised techniques to perform stylized non-factual image captions. Unfortunately, this has barely been explored in unsupervised image captioning methods. Hence, this is a promising area for future work.

Finally, image captioning can be useful in many domains. In some cases, the vocabulary shift between domains is large. Having a single monolithic image captioning dataset that contains vocabulary spanning all domains is impractical; however, creating many different domain-specific image caption datasets is costly and cumbersome. As discussed in previous sections of this paper, unsupervised image captioning methods offer a tool that can help to bridge the gap between different domains. Although some research has been published on this topic [21], [22], [23], most work still depends on domain-specific data in some way. Future work is required to further eliminate this dependence on domain-specific data. Accordingly, this would be an interesting avenue to pursue for future work.

VII. Limitations

This survey has discussed image captioning from the data availability perspective with an emphasis on unsupervised techniques in image captioning. According to the authors best knowledge, the current survey is the first to compare image captioning techniques from this aspect.

Having that said. It is important to note that this survey does not constitute a systematic survey and some of the published work might be unintendedly missed. Additionally, this survey focuses on proposed unsupervised techniques and excludes other areas of image captioning such as the supervised image captioning, datasets, and metrics that are used in measuring models performance. Such areas are critical to the understanding image captioning. Therefore, a comprehensive survey studying all techniques and areas of image caption from various perspective is proposed as a future work.

VIII. Conclusion

In this paper, a survey of semi-supervised and unsupervised image captioning is presented. First existing surveys are examined and compared against the survey presented in this paper. Although previous surveys cover image captioning in details. None of those surveys focuses on unsupervised image captioning. In contrast, this survey puts semi-supervised and unsupervised image captioning techniques under focus. Additionally, the survey categorizes different unsupervised methods based on the data availability and data pairing settings. Where some methods can be used with paired data while others can be used for unpaired data. Furthermore, special cases of unpaired data such as cross-domain and cross lingual image captioning is also discussed. Although such categorization would help to decide which method to use depending on the availability of data. It was not presented in the context of image captioning previously. Finally, the survey presents a discussion on challenges and future research directions of image captioning.

Refrences

[1] Y. Rui, T. Huang and S. Chang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues", Journal of Visual Communication and Image Representation, vol. 10, no. 1, pp. 39-62, 1999. Available: 10.1006/jvci.1999.0413.

[2] M. Hossain, F. Sohel, M. Shiratuddin and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, vol. 51, no. 6, pp. 1-36, 2019. Available: 10.1145/3295748.

[3] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017. Available: 10.1145/3065386.

[4] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1409.0473.

[5] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1411.4555.

[6] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1502.03044.

[7] I. Goodfellow et al., "Generative Adversarial Nets", Advances in Neural Information Processing Systems 27 (NIPS 2014), pp. 2672-2680, 2014.

[8] Y. Feng, L. Ma, W. Liu and J. Luo, "Unsupervised Image Captioning", arXiv.org, 2018. [Online]. Available: https://arxiv.org/abs/1811.10787.

[9] I. Laina, C. Rupprecht and N. Navab, "Towards Unsupervised Image Captioning with Shared Multimodal Embeddings", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1908.09317.

[10] R. Shetty, M. Rohrbach, L. Hendricks, M. Fritz and B. Schiele, "Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training", arXiv.org, 2017. [Online]. Available: https://arxiv.org/abs/1703.10476.

[11] B. Dai, S. Fidler, R. Urtasun and D. Lin, "Towards Diverse and Natural Image Descriptions via a Conditional GAN", arXiv.org, 2017. [Online]. Available: https://arxiv.org/abs/1703.06029.

[12] Y. Wang, J. Xu, Y. Sun and B. He, "Image Captioning based on Deep Learning Methods: A Survey", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1905.08110.

[13] S. Bai and S. An, "A survey on automatic image caption generation", Neurocomputing, vol. 311, pp. 291-304, 2018. Available: 10.1016/j.neucom.2018.05.080.

[14] X. Liu, Q. Xu and N. Wang, "A survey on deep neural network-based image captioning", The Visual Computer, vol. 35, no. 3, pp. 445-470, 2018. Available: 10.1007/s00371-018-1566-y.

[15] R. Bernardi et al., "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures", Journal of Artificial Intelligence Research, vol. 55, pp. 409-442, 2016. Available: 10.1613/jair.4900.

[16] A. Kumar and S. Goel, "A survey of evolution of image captioning techniques", International Journal of Hybrid Intelligent Systems, vol. 14, no. 3, pp. 123-139, 2018. Available: 10.3233/his-170246.

[17] A. Lindh, R. Ross, A. Mahalunkar, G. Salton and J. Kelleher, "Generating Diverse and Meaningful Captions", Artificial Neural Networks and Machine Learning – ICANN 2018, pp. 176-187, 2018. Available: 10.1007/978-3-030-01418-6_18

[18] P. Dognin, I. Melnyk, Y. Mroueh, J. Ross and T. Sercu, "Adversarial Semantic Alignment for Improved Image Captions", the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10463-10471, 2019.

[19] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, "Self-Critical Sequence Training for Image Captioning", in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7008-7024.

[20] T. Chen, Y. Liao, C. Chuang, W. Hsu, J. Fu and M. Sun, "Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner", 2017 IEEE International Conference on Computer Vision (ICCV), pp. 521-530, 2017. Available: 10.1109/iccv.2017.64

[21] W. Zhao et al., "Dual Learning for Cross-domain Image Captioning", Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17, 2017. Available: 10.1145/3132847.3132920

[22] M. Yang et al., "Multitask Learning for Cross-Domain Image Captioning", IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1047-1061, 2019. Available: 10.1109/tmm.2018.2869276

[23] L. Wang, A. Schwing and S. Lazebnik, "Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space", 2017. [Online]. Available: https://arxiv.org/abs/1711.07068.

[24] Y. Xia et al., "Dual Learning for Machine Translation", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1611.00179.

[25] J. Gu, S. Joty, J. Cai and G. Wang, "Unpaired Image Captioning by Language Pivoting", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1803.05526.

[26] Y. Song, S. Chen, Y. Zhao and Q. Jin, "Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1908.05407.

[27] W. Lan, X. Li and J. Dong, "Fluency-Guided Cross-Lingual Image Captioning", Proceedings of the 2017 ACM on Multimedia Conference - MM '17, 2017. Available: 10.1145/3123266.3123366

[28] X. Liu, H. Li, J. Shao, D. Chen and X. Wang, "Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1803.08314.

[29] D. Kim, J. Choi, T. Oh and I. Kweon, "Image Captioning with Very Scarce Supervised Data: Adversarial Semi-Supervised Learning Approach", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1909.02201v1.

[30] P. Anderson, S. Gould and M. Johnson, "Partially-Supervised Image Captioning", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1806.06004

[31] J. Gu, S. R. Joty, J. Cai, H. Zhao, X. Yang and G. Wang, "Unpaired image captioning via scene graph alignments", 2019. [Online]. Available: https://arxiv.org/abs/1903.10658

[32] T. N. Kipf and M. Welling, "Variational graph auto-encoders", Proc. NIPS Workshop Bayesian Deep Learn., pp. 1-3, 2016.

[33] D. Guo, Y. Wang, P. Song and M. Wang, "Recurrent relational memory network for unsupervised image captioning", Proc. 29th Int. Joint Conf. Artif. Intell., pp. 920-926, Jul. 2020.

[34] T.-Y. Lin et al., "Microsoft COCO: Common objects in context", *Proc. Eur. Conf. Comput. Vis.*, pp. 740-755, 2014.

[35] Maria-Elena Nilsback and Andrew Zisserman, "Automated Flower Classification over a Large Number of Classes", *ICVGIP 2008*, pp. 722-729, 2008.

[36] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein GAN" in arXiv:1701.07875, 2017, [online] Available: http://arxiv.org/abs/1701.07875