Hindawi
Complexity
Volume 2021, Article ID 5538927, 19 pages
https://doi.org/10.1155/2021/5538927

Research Article

Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning

Ariyo Oluwasammi,[1] **Muhammad Umar Aftab**,[2] **Zhiguang Qin**,[1] **Son Tung Ngo**,[3] **Thang Van Doan**,[3] **Son Ba Nguyen**,[3] **Son Hoang Nguyen**,[3] **and Giang Hoang Nguyen**[3]

[1]*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2]*Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Chiniot-Faisalabad Campus, Chiniot 35400, Pakistan*
[3]*ICT Department, FPT University, Hanoi 10000, Vietnam*

Correspondence should be addressed to Zhiguang Qin; qinz@uestc.edu.cn

Received 8 January 2021; Revised 31 January 2021; Accepted 6 March 2021; Published 23 March 2021

Academic Editor: Dan Selisteanu

With the emergence of deep learning, computer vision has witnessed extensive advancement and has seen immense applications in multiple domains. Specifically, image captioning has become an attractive focal direction for most machine learning experts, which includes the prerequisite of object identification, location, and semantic understanding. In this paper, semantic segmentation and image captioning are comprehensively investigated based on traditional and state-of-the-art methodologies. In this survey, we deliberate on the use of deep learning techniques on the segmentation analysis of both 2D and 3D images using a fully convolutional network and other high-level hierarchical feature extraction methods. First, each domain's preliminaries and concept are described, and then semantic segmentation is discussed alongside its relevant features, available datasets, and evaluation criteria. Also, the semantic information capturing of objects and their attributes is presented in relation to their annotation generation. Finally, analysis of the existing methods, their contributions, and relevance are highlighted, informing the importance of these methods and illuminating a possible research continuation for the application of semantic image segmentation and image captioning approaches.

1. Introduction

The data of optical perception are becoming increasingly available in large volume nowadays, creating a crucial use in several real-world applications such as quality assurance, medical analysis, surveillance, autonomous vehicles, face recognition, forensic and biometrics, and 3D reconstruction [1–4]. This upsurge in the bulk of digital images and video has directed the creation of computer vision (CV), a branch of computer science (CS). From a general overview, computer vision relates to the use of the computer to gain a high-level understanding of images and videos [5]. Rather than manual operations, it encompasses the automatic acquisition, processing, and analyzing of large data for the sole purpose of extracting patterns and intuition. In most cases, computer vision seeks to apply artificial intelligence (AI) theory, equations, tools, frameworks, and algorithms for accomplishing the task of helping computers to see and also understand the content of both digital and analog world through the mimicking of the human visual system [6]. Although seeing and understanding seems a trivial task or very easy for humans, it is nevertheless a complex problem for computers partly because of our limited understanding of how the human brain works and how it processes things [7]. However, through years of research and technological advancement, some feats have been achieved, and computer vision has extensively evolved [8–11]. Today, semantic segmentation remains a huge challenge in the scope of image

and video understanding alongside image captioning which combines computer vision with another branch of artificial intelligence called natural language processing (NLP) to derive sentence description of an image [12]. Notwithstanding, as with all other AI-related tasks, a modern subset of machine learning (ML), namely, deep learning (DL), has been the evolution of machine learning, producing state-of-the-art results in almost all of the tasks compared to other traditional algorithms such as decision trees, naive Bayes, support vector machines (SVMs), ensembles, and clustering algorithms [13–16].

Deep learning, as a branch of machine learning, uses layers of artificial neural networks to imitate the human neural networks in decoding intuition from a large amount of data automatically [17] and is unlike other machine learning algorithms which rely heavily on feature engineering, utilizing domain knowledge in the creation of feature extractors [18]. The stacked layer of neural networks represents feature hierarchy as simple features at the initial layers are reconstructed from one layer to another in forming complex features [19]. As a result, the deeper networks are computationally intensive to model and train, leading to the manufacture of more advanced computational chips, including Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) [20, 21]. Presently, several deep learning models exist, and some of the most popular ones include recurrent neural networks (RNNs), autoencoders, convolutional neural network (CNN), deep belief networks (DBNs), and deep Boltzmann machine (DBM) [22–25]. Among the most common deep learning algorithms, the convolutional neural network is the most suitable for analyzing visual imagery because of its shift and space invariant characteristics, taking advantage of hierarchical learning in combining simpler patterns to form complex patterns and structures [26]. Using the shared weights architectural pattern of filters [27–29], each filter represents different features of the input data which when summed can yield more complex structures [30–32].

In this paper, our prime motivation is fixating on the recent deep learning segmentation techniques of both 2D and 3D images using fully convolutional network and other high-level hierarchical feature extraction methods as an integral component of computer vision. This is further expanded into the generation of captions for images, emerging as a subset of artificial intelligence's natural language processing. Furthermore, we review the discussed models' accomplishments by comparing their evaluation, which indicates the most effective and efficient approaches for different tasks and challenges encountered. This, we believe, is enlightening as it provides insight for the further evolution of practical model design.

This paper is organized as follows: it introduces segmentation, popular segmentation algorithms, characteristics, datasets, and evaluation in Section 2. An introduction of captioning and its various models are followed in Section 3 alongside available datasets, evaluation metrics, and a comparative discussion of the models. Finally, Section 4 concludes the paper with the overall summary of the typical problems, solutions, and possible directions in semantic segmentation and image captioning.

2. Semantic Segmentation

Semantic segmentation relates to the process of pixel-level classification of images such that each pixel in an image is classified into a distinguished class cluster [33]. Since the inception of deep learning, semantic segmentation has been a pivotal area of image processing and computer vision which has seen major research and application in several domains [34]. Image segmentation recognizes boundaries between objects in an image by using line and curve segments to categorize such objects, while instance segmentation, however, classifies instances of all the available classes in an image such that all objects are identified as a separate entity. All the same, semantic segmentation differs from ordinary segmentation which, on the one hand, only expresses the partitioning of an image into clusters without a tangible intuitive attempt at understanding the partitioned clusters or relating them with one another [35]. Semantic segmentation, on the other hand, as the name implies, tries to describe semantically meaningful objects in an image based on their well-defined association and understanding [36]; these differences are well depicted in Figure 1.

2.1. Methods and Approaches

2.1.1. Traditional Methods. During the pre-ANN era, most segmentation and semantic segmentation approaches were predominantly thresholding and clustering algorithms which are largely unsupervised methods. In most cases, traditional semantic segmentation methods consume less time for model computation. Also, most of the approaches require less data than the modern-day era of artificial neural networks and deep learning. The simplest, by all means, is the thresholding techniques which apply pixel intensity as the criteria for distribution. For binary segmentation, a single threshold value is required, and pixels on both sides of the threshold are classified separately into two distinct classes. There are also advance forms of thresholding involving more classes, and they are often grouped as histogram shape-based, entropy-based, object attribute-based, and spatial-based [37].

K-means clustering uses a predefined number of centroids to determine the number of clusters in which objects are to be categorized. The centroids are randomly selected at the beginning and then are iteratively adjusted by computing the distance apart from other points in the dataset, assigning each point to the closest centroid [38]. Fuzzy C-means (FCM), a technically advanced form of *K*-means, allows classification of data points into many label classes based on the level of membership [39]. This is of advantage in situations where dataset texture overlaps or does not have a well-defined cluster [40]. Gaussian mixture model (GMM) is also often used for both hard clustering and soft clustering by assigning the pixel to the component having the highest posterior probability [41]. GMM assumes that the data's Gaussian distributions represent the number of clusters

(a) (b) (c)

Figure 1: Illustration of differences in segmentation: (a) object detection, (b) semantic segmentation, and (c) instance segmentation [36].

available in the data, and it uses the expectation-maximization (EM) algorithm to determine missing latent variables [42]. Random forest [43], naive Bayes [44], ensemble modeling [45], Markov random network (MRF) [46], and support vector machines (SVMs) [47] are also techniques that are useful for several tasks, especially classification and regression [48].

2.1.2. Region-Based Models. In the region-based semantic segmentation design, regions are first extracted in an image and described based on their constituent features [49]. Then, a region classifier that has been trained is used to label pixels per region with which it has the highest occurrence. The region-based approaches use the divide and conquer method such that many features are captured using multiscale features and then combined to form a whole. In cases where objects overlap on several regions, the classifier either determines the most suitable region or the model is set to select the region with the maximum value [50]. This often causes pixel mismatch, but a postprocessing operation is mostly used to reduce the effects [51].

Region CNN (R-CNN) uses a bounding box to identify and classify objects in an image by proposing several boxes convolving an image and identifying if they correspond to an object [52, 53]. The process of selective search is used in creating boundary boxes of varying window sizes for region proposal, and each of the boxes classifies objects based on different properties, making the algorithm quite impressive but slow [54]. To overcome the drawbacks of the R-CNN, Fast R-CNN [55] was proposed which eliminates the redundancy in the proposed region, thereby lessening the computational requirements. The R-CNN model was replaced with a single CNN per image whose computation would be shared among the proposals, using the region of interest pooling technique and training all the models including the use of convolutional neural network to classify the images, and bounding boxes regressor as a single entity. The Faster R-CNN [56] uses a region proposal network (RPN) to obtain a proposal from the network instead, while the Mask R-CNN [57] was extended to include a pixel-level segmentation. Technically, the Mask R-CNN replaces the region of interest pooling module in the Faster R-CNN with another which has an accurate alignment module. Also, it includes an additional parallel branch for segmentation mask prediction [58].

2.1.3. Fully Convolutional Network-Based Models. Fully convolutional network (FCN) models do not have dense layers, such as in other traditional CNNs; they are composed of 1×1 convolutions that achieve the task of dense layers or fully connected layers. Also, fully convolutional network (FCN), as displayed in Figure 2, takes images of arbitrary sizes as the input and returns outputs of corresponding spatial dimensions. This model principally builds on the encoder-decoder model to classify pixels in an image into predefined classes by using a convolution network in the encoder to extract features, thereby reducing the feature maps' dimensionality before being upsampled by the decoder (SegNet) [60]. During convolutional neural network computation, input images are downsized, resulting in a smaller output with reduced spatial features. This problem is solved via the upsample technique, which transposes the downsized images to a larger size, making pixelwise comparison efficient and effective. Some upsampling methods such as transpose convolutions are learned, thus increasing model complexity and computation, while several others exclude learning including nearest neighbor, the bed of nails, and max unpooling [61]. The fully convolutional network is majorly trained as an end-to-end model to compute pixelwise loss and trained using the backpropagation approach. The FCN was firstly inspired by Long et al. [59] using the popular AlexNet CNN architecture in the encoder and transpose convolution layers in the decoder to upsample the feature to the desired dimension. A variant FCN having skip connections from previous layers in the network was proposed named UNet [62]. UNet intends to compliment the learned features with fine-grain details from contracting paths to capture the context and enhance classification accuracy.

Residual blocks were introduced with shorter skip connections between the encoder and decoder, granting faster convergence of deeper models during training [63]. Multiscale induction in the dense blocks conveys low-level features across layers to ones with high-level features, resulting in feature reuse [64]. This makes the model easier to train and improve the performance as well. An architecture having two-stream CNN uses its gates to process the image shape in different branches, then connected and fused at a later stage. The model also proposed a new loss function that aligns segmentation prediction with the ground truth boundaries [65]. An end-to-end single-pass model applying dense decoder shortcut connections extracts semantics from

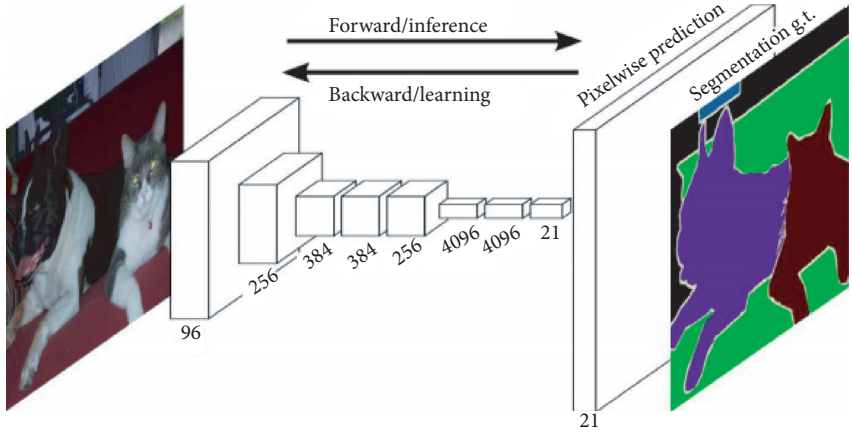

Figure 2: Illustration of fully convolutional networks for semantic segmentation [59].

high-level features such that propagation of information from one block to another combines multiple-level features [66]. The model designed based on the ResNeXt's residual building blocks helps it to aggregate blocks of feature captures which are fused as output resolutions [67].

ExFuse aims to connect the gap between low- and high-level features in convolutional networks by introducing semantic information at the lower-level features as well as high resolutions into the high-level features. This was achieved by proposing two fusion methods named explicit channel resolution embedding and densely adjacent prediction [68]. Contrary to most models, a balance between model accuracy and speed was achieved in ICNet which consolidates several multiresolution branches by introducing an image cascade network that allows real-time inference. The network cascade image inputs into different segments as low, medium, and high resolution before being trained with this label guidance [69].

2.1.4. Refinement Network. Because of the resolution reduction caused by the encoder models in the typical FCN-based model, the decoder has inherent problems of producing fine-grained segmentation, especially at the boundaries and edges [70]. Though this problem has been tackled by incorporating skip connections, adding global information, and others means, the problem is by no means solved, and some algorithms have involved several features or, in some cases, certain postprocessing functions to find alternative solutions [71]. DeepLab1 [72] combines ideas from the deep convolutional neural network and probabilistic graphical models to achieve pixel-level classification. The localization problem of the neural network output layer was remedied using a fully connected conditional random field (CRF) as a means of performing segmentation with controlled signal extermination. DeepLab1 applied atrous convolutions instead of the regular convolutions which accomplish the learning of aggregate multiscale contextual features. Visible in Figure 3, DeepLab1 allows the expansion of kernel window sizes without increasing the number of weights. The multiscale atrous convolutions help to overcome the problem of insensitivity to fine details by other models and decrease output blurriness. This could result in additional complexity in computation and time depending on the postprocessing network's computational processes.

The ResNet deep convolutional network architecture was applied in DeepLab2 which enables the training of various distinct layers while preserving the performance [73]. Besides, DeepLab2 uses atrous spatial pyramid pooling (ASPP) to capture long-range context. The existence of objects at different scales and the reduced feature resolution problems of semantic segmentation are tackled by designing a cascade of atrous convolutions which could run in parallel to capture various scales of context information alongside global average pooling (GAP) to embed context information features [74]. FastFCN implements joint pyramid upsampling which substitutes atrous convolutions to free up memory and lessen computations. Using a fully connected network framework, the joint pyramid upsampling technique extracts feature maps of high resolution into a joint upsampling problem. The models used atrous spatial pyramid pooling to extract the last three-layer output features and a global context module to map out the final predictions [75]. The atrous spatial pyramid pooling limitation of lack of dense feature resolution scale is attempted by concatenating multiple branches of atrous-convolved features at different rates which are later fused into a final representation, resulting in dense multiscale features [76].

2.1.5. Weakly Supervised and Semisupervised Approaches. Though most models depend on a large number of images and their annotated label, the process of manually annotating labels is quite daunting and time-consuming, so semantic segmentation models have been attempted with weakly supervised approaches. Given weakly annotated data at the image level, the model was trained to assign higher weights to pixels corresponding with the class label. Trained on a subset of the ImageNet dataset, during training, the networks focus on recognizing important pixels relating to a prior labeled single-class object and matching them to the class through inference [77]. Bounding box annotation is used to train semantic labeling of image segmentation which accomplishes 95% quality of fully supervised models. The

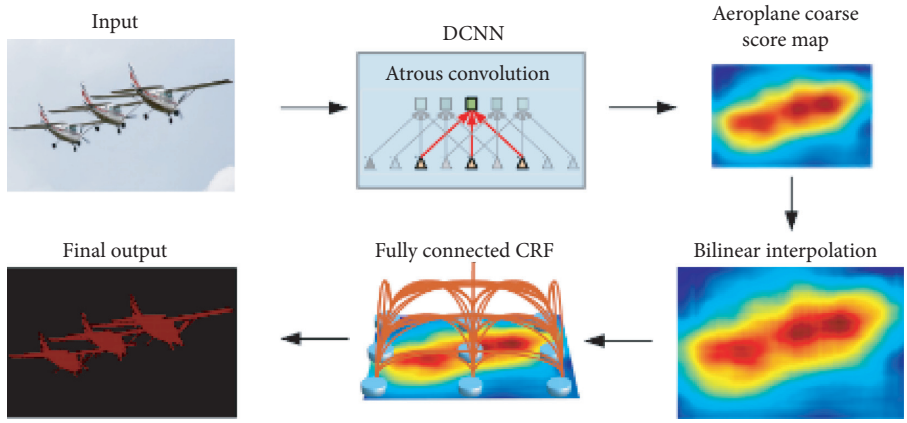

Figure 3: DeepLab framework with fully connected CRFs [72].

bounding box and information of the constituent object were used prior to training [78]. A model combining both labeled and weakly annotated images with a clue of the presence or absence of a semantic class was developed using the deep convolutional neural network and expectation-maximization (EM) algorithm [79].

BoxSup iteratively generates automatic region proposals while training convolutional networks to obtain segmentation masks and as well as improve the model's ability to classify objects in an image. The network uses bounding box annotations as a substitute for supervised learning such that regions are proposed during training to determine candidate masks, overtime improving the confidence of the segmentation masks [80]. A variant of generative adversarial learning approach which constitutes a generator and discriminator was used to design a semisupervised semantic segmentation model. The model was first trained using full labeled data which enable the model's generator to learn the domain sample space of the dataset which is leveraged to supervise unlabeled data. Alongside the cross-entropy loss, an adversarial loss was proposed to optimize the objective function of tutoring the generator to generate images as close as possible to the image labels [81].

2.2. Datasets. Deep learning requires an extensive amount of training data to comprehensively learn patterns and fine-tune the number of parameters needed for its gradient convergence. Accordingly, there are several available datasets specifically designed for the task of semantic segmentation which are as follows:

PASCAL VOC: PASCAL Visual Object Classes (VOC) [82] is arguably the most popular semantic segmentation dataset with 21 classes of predefined object labels, background included. The dataset contains images and annotations which could be used for detection, classification, action classification, person layout, and segmentation tasks. The dataset's training, validation, and test set has 1464, 1449, and 1456 images, respectively. Yearly, the dataset has been used for public competitions since 2005.

MS COCO: Microsoft Common Objects in Context [83] was created to push the computer vision state of the art with standard images, annotation, and evaluation. Its object detection task dataset uses either instance/object annotated features or a bounding box. In total, it has 80 object categories with over 800,000 available images for its training, test, and validation sets, as well as over 500,000 object instances that are segmented.

Cityscapes: Cityscapes dataset [84] has a huge amount of images taken from 50 cites during different seasons and times of the year. It was initially a video recording, and the frames were extracted as images. It has 30 label classes in about 5000 densely annotated images and 20,000 weakly annotated images which have been categorized into 8 groups of humans, vehicles, flat surfaces, constructions, objects, void, nature, and sky. It was primarily designed for urban scene segmentation and understanding.

ADE20K: ADE20K dataset [85] has 20,210 training images, 2000 validation images, and 3000 test images which are well suited for scene parsing and object detection. Alongside the 3-channel images, the dataset contains segmentation masks, part segmentation masks, and a text file that contains information about the object classes, identification of instances of the same class, and the description of each image's content.

CamVid: CamVid [86] is also a video sequence of scenes which have been extracted into images of high resolution for segmentation tasks. It consists of 101 images of $960 * 720$ dimension and their annotations which have been classified into 32 object classes including void, indicating areas which do not belong to a proper object class. The dataset RGB class values are also available, ranging from 0 to 255.

KITTI: KITTI [87] is popularly used for robotics and autonomous car training, focusing extensively on 3D tracking, stereo, optical flow, 3D object detection, and visual odometry. The images were obtained through two high-resolution cameras attached to a car driving around the city of Karlsruhe, Germany, while their annotations were done by a Velodyne laser scanner. The data aim to reduce bias in existing benchmarks with a standard evaluation metric and website.

SYNTHIA: SYNTHetic [88] Collection of Imagery and Annotations (SYNTHIA) is a compilation of imaginary images from a virtual city that has a high pixel-level

resolution. The dataset has 13 predefined label classes consisting of road, sidewalk, fence, sky, building, sign, pedestrian, vegetation, pole, and car. It has a total of 13,407 training images.

2.3. Evaluation Metrics. The performance of segmentation models is computed mostly in the supervised scope whereby the model's prediction is compared with the ground truth at the pixel level. The common evaluation metrics are pixel accuracy (PA) and intersection over union (IoU). Pixel accuracy refers to the ratio of all the pixels classified in their correct classes to the total amount of pixels in the image. Pixel accuracy is trivial and suffers from class imbalance such that certain classes immensely dominate other classes.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{PA} &= \frac{\sum_i n_{ii}}{\sum_i t_i}, \end{aligned} \tag{1}$$

where n_c is represented as the number of classes and n_{ii} is also represented as the number of pixels of class i which are predicted to class i, while n_{ij} represents the number of pixels of class i which are predicted as class j with the total number of pixels of a particular class i represented as $t_i = \sum_j n_{ij}$.

Mean pixel accuracy (mPA) improves pixel accuracy slightly; it computes the accuracy of the images per class instead of a global computation of all the classes. The mean of the class accuracies is then computed to the overall number of classes.

$$\text{mPA} = \frac{1}{n_c} \sum_i \frac{n_{ii}}{t_i}. \tag{2}$$

Intersection over union (IoU) metric, which is also referred to as the Jaccard index, measures the percentage overlap of the ground truth to the model prediction at the pixel level, thereby computing the amount of pixels common with the ground truth label and mask prediction [89].

$$\text{mIoU} = \frac{1}{n_c} \frac{\sum_i n_{ii}}{\sum_j n_{ij} + \sum_j n_{ji} - n_{ii}}. \tag{3}$$

2.4. Discussion. Different machine learning and deep learning algorithms and backbones yield different results based on the models' ability to learn mappings from input images to the label. In tasks involving images, CNN-based approaches are by far the most expedient. Although they can be computationally expensive compared to other simpler models, such models occupy a bulk of the present state of the art. Traditional machine learning algorithms such as random forest, naive Bayes, ensemble modeling, Markov random field (MRF), and support vector machines (SVMs) are too simple and rely heavily on domain feature understanding or handcrafted feature engineering, and in some cases, they are not easy to fine-tune. Also, clustering algorithms such as K-means and fuzzy C-means mostly require that the number of clusters is specified beforehand, and they are not very effective with multiple boundaries.

Because of the CNN's invariant property, it is very effective for spatial data and object detection and localization. Besides, the modern backbone of the fully convolutional network has informed several methods of improving segmentation localization. First, the decoder uses upsampling techniques to increase the features' resolution, and then skip connections are added to achieve the transfer of fine-grain features to the other layers. Furthermore, postprocessing operations as well as the context and attention networks have been exploited.

The supervised learning approach still remains the dormant technique as there have been many options for generating datasets as displayed in Table 1. Data augmentation involving operations such as scaling, flipping, rotating, scaling, cropping, and translating has made multiplication of data possible. Also, the application of the generative adversarial network (GAN) has played a major role in the replication of images and annotations.

3. Image Captioning

Image captioning relates to the general idea of automatically generating the textual description of an image, and it is often also referred to as image annotation. It involves the application of both computer vision and natural language processing tools to achieve the transformation of imagery depiction into a textual composition [111]. Tasks such as captioning were almost impossible prior to the advent of deep learning, and with advances in sophisticated algorithms, multimodal techniques, efficient hardware, and a large bulk of datasets, such tasks are becoming easy to accomplish [112]. Image captioning has several applications to solving some real-world problems including providing aid to the blind, autonomous cars, academic bot, and military purposes. To a large extent, the majority of image captioning success so far has been from the supervised domain whereby huge amounts of data consisting of images and about two to five label captions describing the actions of the images are provided [113]. This way, the network is tasked with learning the images' feature presentation and mapping them to a language model such that the end goal of a captioning network is to generate a textual representation of an image's depiction [114].

Though characterizing an image in the form of text seems trivial and straightforward for humans, it is by no means simple to be replicated in an artificial system and requires advanced techniques to extract the features from the images as well as map the features to the corresponding language model. Generally, a convolutional neural network (CNN) is tasked with feature extraction, while a recurrent neural network (RNN) relies upon to translate the training annotations with the image features [115]. Aside from determining and extracting salient and intricate details in an image, it is equally important to extract the interactions and semantic relationship between such objects and how to illustrate them in the right manner using appropriate tenses and sentence structures [116]. Also, because the training

Table 1: Class pixel label distribution in the CamVid dataset.

Dataset	Method	mIoU
CamVid	ApesNet [90]	48.0
	ENet [91]	51.3
	SegNet [60]	55.6
	LinkNet [92]	55.8
	FCN8 [59]	57.0
	AttentionM [93]	60.1
	DeepLab-LFOV [72]	61.6
	Dilation8 [66]	65.3
	BiseNet [94]	68.7
	PSPNet [60]	69.1
	DenseDecoder [67]	70.9
	AGNet [95]	75.2
PASCAL VOC	Wails [96]	55.9
	FCN8 [59]	62.2
	PSP-CRF [97]	65.4
	Zoom Out [98]	69.6
	DCU [99]	71.7
	DeepLab1 [72]	71.6
	DeConvNet [61]	72.5
	GCRF [100]	73.2
	DPN [101]	74.1
	Piecewise [102]	75.3
Cityscapes	FCN8 [59]	65.3
	DPN [101]	66.8
	Dilation10 [103]	67.1
	LRR [104]	69.7
	DeepLab2 [73]	70.4
	FRRN [105]	71.8
	RefineNet [106]	73.6
	GEM [107]	73.69
	PEARL [108]	75.4
	TuSimple [109]	77.6
	PSPNet [110]	78.4
	SPP-DCU [99]	78.9

labels which are texts are different from the features obtained from the images, language model techniques are required to analyze the form, meaning, and context of a sequence of words. This becomes even more complex as keywords are required to be identified for emphasizing the action or scene being described [117].

Visual features: deep convolutional neural network (DCNN) is often used as the feature extractor for images and videos because of the CNN's invariance property [118] such that it is able to recognize objects regardless of variation in appearances such as size, illumination, translation, or rotation as displayed in Figure 4. The distortion in pixel arrangement has less impact on the architecture's ability to learn essential patterns in the identification and localization of the crucial features. Essential feature extraction is paramount, and this is easily achieved via the CNN's operation of convolving filters over images, subsequently generating feature maps from the receptive fields from which the filters are applied. Using backpropagation techniques, the filter weights are updated to minimize the loss of the model's prediction compared to the ground truth [119]. There have been several evolutions over the years, and this has ushered considerate architectural development in the extraction methodology. Recently, the use of a pretrained model has been explored with the advantage of reducing time and computational cost while preserving efficiency. These extracted features are passed along to other models such as the language decoder in the visual space-based methods or to a shared model as in the multimodal space for image captioning training [120].

Captioning: image caption or annotation is an independent scope of artificial intelligence, and it mostly combines two models consisting of a feature extractor as the encoder and a recurrent decoder model. While the extractor obtains salient features in the images, the decoder model which is similar in pattern to the language model utilizes a recurrent neural network to learn sequential information [121]. Most captioning tasks are undertaken in a supervised manner whereby the image features act as the input which are learned and mapped to a textual label [122]. The label captions are first transformed into a word vector and are combined with the feature vector to generate a new textual description. Most captioning architectures follow the partial caption technique whereby part of the label vector is combined with the image vector to predict the next word in the sequence [123]. Then, the prior words are all combined to predict the next word and continued till an end token is reached. In most cases, to infuse semantics and intuitive representation into the label vector, a pretrained word embedding is used to map the dimensional representation of the embeddings into the word vector, enriching its content and generalization [124].

3.1. Image Captioning Techniques

3.1.1. Retrieval-Based Captioning. Early works in image captioning were based on caption retrieval. Using this technique, the caption to a target image is generated by retrieving the descriptive sentence of such an image from a set of predefined caption databases. In some cases, the newly generated caption would be one of the existing retrieved sentences or, in other cases, could be made up of several existing retrieved sentences [125]. Initially, the features of an image are compared to the available candidate captions or achieved by tagging the image property in a query. Certain properties such as color, texture, shape, and size were used for similarity computation between a target image and predefined images [126]. Captioning via the retrieval method can be retrieved through the visual and multimodal space, and these approaches produce good results generally but are overdependent on the predefined captions [127].

Specific details such as the object in an image or the action or scene depicted were used to connect images to the corresponding captions. This was computed by finding the ratio of similarity between such information to the available sentences [128]. Using the kernel canonical correlation analysis technique, images and related sentences were ranked based on their cosine similarities after which the most similar ones were selected as the suitable labels [129], while the image features were used for reranking the ratio of image-text correlation [130]. The edges and contours of images were utilized to obtain

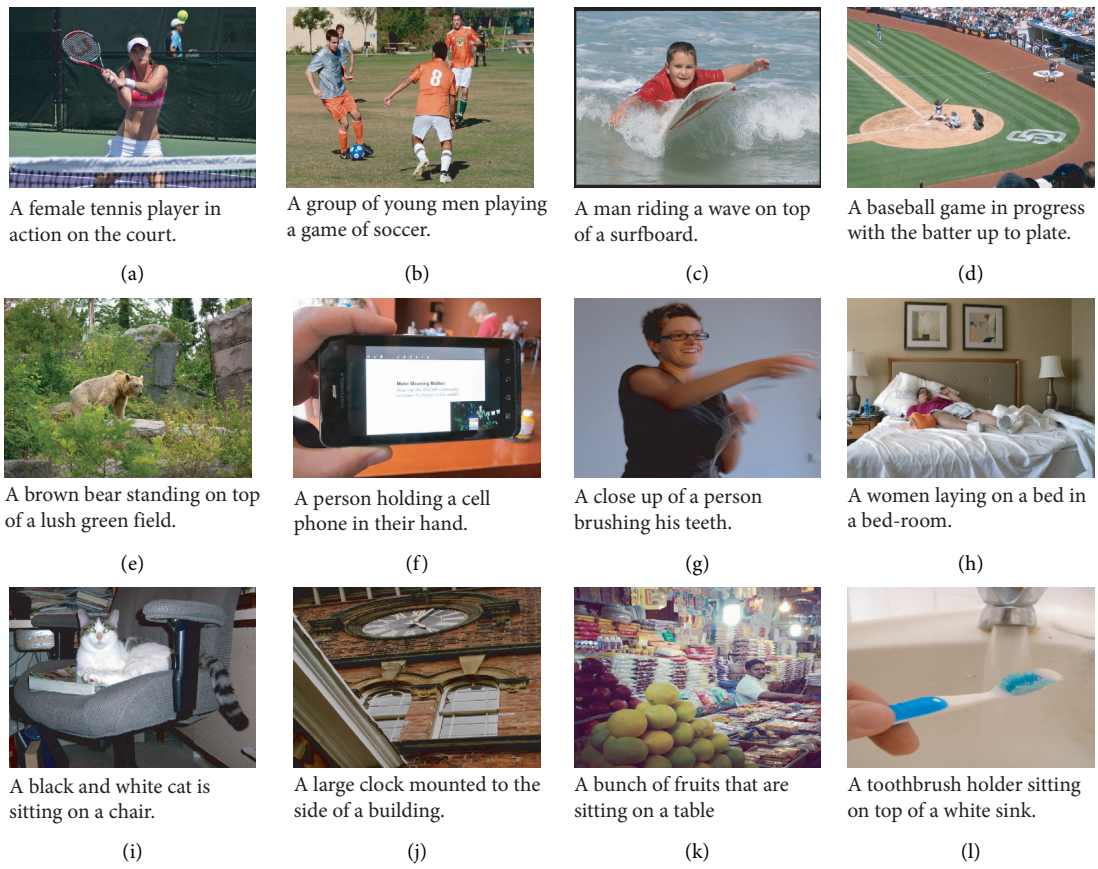

Figure 4: Sample images and their corresponding captions [112].

pattern and sketches such that they were used as a query for image retrieval, whereby the generated sketches and the original images were structured into a database [131]. Building on the logic of the original image and its sketch, more images were dynamically included alongside their sketches to enhance learning [132]. Furthermore, deep learning models were applied to retrieval-based captioning by using convolutional neural networks (CNNs) to extract features from different regions in an image [133].

3.1.2. Template-Based Captioning. Another common approach of image annotation is template-based which involves the identification of certain attributes such as object type, shape, actions, and scenes in an image, which are then used in forming sentences in a prestructured template [134]. In this method, the predetermined template has a constant number of slots such that all the detected attributes are then positioned in the slots to make up a sentence. In this case, the words representing the detected features make up the caption and are arranged such that they are syntactically and semantically related, thus generating grammatically correct representations [135]. The fixed template problem of the template-based approach was overcome by incorporating a parsed language model [136], giving the network higher flexibility and ability to generate better captions.

The underlying nouns, scenes, verbs, and prepositions determining the main idea of a sentence were explored and trained using a language model to obtain the probability distribution of such parameters [137]. Certain human postures and orientation which do not involve the movement of the hands such as walking, standing, and seeing and the position of the head were used to generate captions of an image [138]. Furthermore, the postures were extended to describe human behavior and interactions by incorporating motion features and associating them with the corresponding action [139]. Each body part and the action it is undergoing are identified, and then this is compiled and integrated to form a description of the complete human body. Human posture, position, and direction of the head and position of the hands were selected as geometric information for network modeling.

3.1.3. Neural Network-Based Captioning. Compared to other machine learning algorithms or preexisting approaches, deep learning has achieved astonishing heights in image captioning, setting new benchmarks with almost all of the datasets in all of the evaluation metrics. These deep learning approaches are mostly in the supervised setting which requires a huge amount of training data including both images and their corresponding caption labels. Several models have been applied such as artificial neural network (ANN), convolutional neural network (CNN), recurrent neural network (RNN), autoencoder, generative adversarial network (GAN), and even a combination of one or more of them.

Dense captioning: dense captioning emerges as a branch of computer vision whereby pictorial features are densely annotated depending on the object, object's motion, and its interaction. The concept depends on the identification of features as well as their localization and finally expressing such features with short descriptive phrases [140]. This idea is drawn from the understanding that providing a single description for a whole picture can be complex or sometimes bias; therefore, a couple of annotations are generated relating to different recognized salient features of the image. The training data of a dense caption in comparison to a global caption are different in that various labels are given for individual features identified by bounding boxes, whereby a general description is given in the global captioning without a need for placing bounding boxes on the images [141]. Visible in Figure 5, a phrase is generated from each region in the image, and these regions could be compiled to form a complete caption of the image. Generally, dense captioning models face a huge challenge as most of the target regions in the images overlap which makes accurate localization challenging and daunting [143].

The intermodal alignment of the text and images was investigated on region-level annotations which pioneers a new approach for captioning, leveraging the alignment between the feature embedding and the word vector semantic embedding [144]. A fully convolutional localization network (FCLN) was developed to determine important regions of interest in an image. The model combines a recurrent neural network language model and a convolutional neural network to enforce the logic of object detection and image description. The designed network uses dense localization layer and convolution anchors built on the Faster R-CNN technique to predict region proposal from the input features [142]. A contextual information model that combines previous and future features of a video spanning up to two minutes achieved dense captioning by transforming the video input into slices of frames. With this, an event proposal module helps to extract the context from the frames which are fed into a long short-term memory (LSTM) unit, enabling it to generate different proposals at different time steps [145]. Also, a framework having separate detection network and localization captioning network accomplished improved dense captioning with faster speed by directly producing detected features rather than via the common use of the region proposal network (RPN) [146].

Encoder-decoder framework: most image captioning tasks are built on the encoder-decoder structure whereby the images and texts are managed as separate entities by different models. In most cases, a convolutional neural network is presented as the encoder which acts as a feature extractor, while a recurrent neural network is presented as the decoder which serves as a language model to process the extracted features in parallel with the text label, consequently generating predicted captions for the input images [147]. CNN helps to identify and localize objects and their interaction, and then this insight is combined with long-term dependencies from a recurrent network cell to predict a word at a time, depending on the image context vector and previously generated words [148]. Multiple CNN-based encoders were proposed to provide a more comprehensive and robust capturing of objects and their interaction from images. The idea of applying multiple encoders is suggested to complement each unit of the encoder to obtain better feature extraction. These interactions are then translated to a novel recurrent fusion network (RFNet) which could fuse and embed the semantics from the multiple encoders to generate meaningful textual representations and descriptions [149].

Laid out in Figure 6, a combination of two CNN models as both encoder and decoder was explored to speed up computational time for image captioning tasks. Because RNN's long-range information is computed step by step, this causes very expensive computation and is solved by stacking layers of convolution to mimic tree structure learning of the sentences [150]. Three distinct levels of features which are regional, visual, and semantic features were encoded in a model to represent different analyses of the input image, and then this is fed into a two-layer LSTM decoder for generating well-defined captions [151]. A concept-based sentence reranking technique was incorporated into the CNN-LSTM model such that concept detectors are added to the underlying sentence generation model for better image description with minimal manual annotation [152]. Furthermore, the generative adversarial network (GAN) was conditioned on a binary vector for captioning. The binary vector represented some form of sentiment which the image portrays and then was used to train the adversarial model. The model took both images and an adjective or adjective-noun pair as the input to determine if the network could generate a caption describing the intended sentimental stance [153].

Attention-guided captioning: attention has become increasingly paramount, and it has driven better benchmarks in several tasks such as machine translation, language modeling, and other natural language processing tasks, as well as computer vision tasks. In fact, attention has proven to correlate the meaning between features, and this helps to understand how such a feature relates to one another [154]. Incorporating this into a neural network, it encourages the model to focus on salient and relevant features and pay less consideration to other noisy aspects of the data space distribution [155]. To estimate the concept of attention in image annotation, a model is trained to concentrate its computation on the identified salient regions while generating captions using both soft and hard attention [156]. The deterministic soft attention which is trainable via standard backpropagation is learned by weighting the annotated vector of the image features, while the stochastic hard attention is trained via maximizing a variational lower bound, setting it to 1 when the feature is salient [157].

Following the where and what analysis of what the model should concentrate on, adaptive attention used a hierarchical structure to fuse both high-level semantic information and visual information from an image to form intuitive representation [120]. The top-down and bottom-up approaches are fused using semantic attention which first defines attribute detectors that dynamically enable it to switch between concepts. This empowers the detectors to determine suitable candidates for attention computation based on the specified

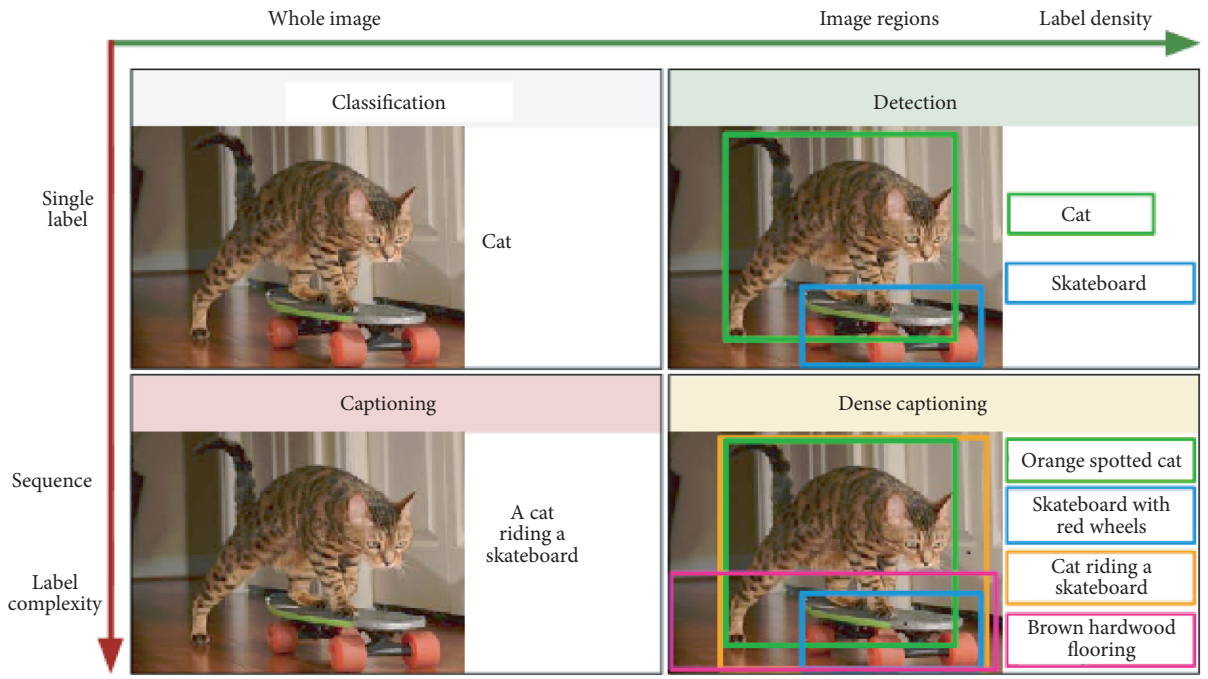

Figure 5: Dense captioning illustrating multiple annotations with a single forward pass [142].

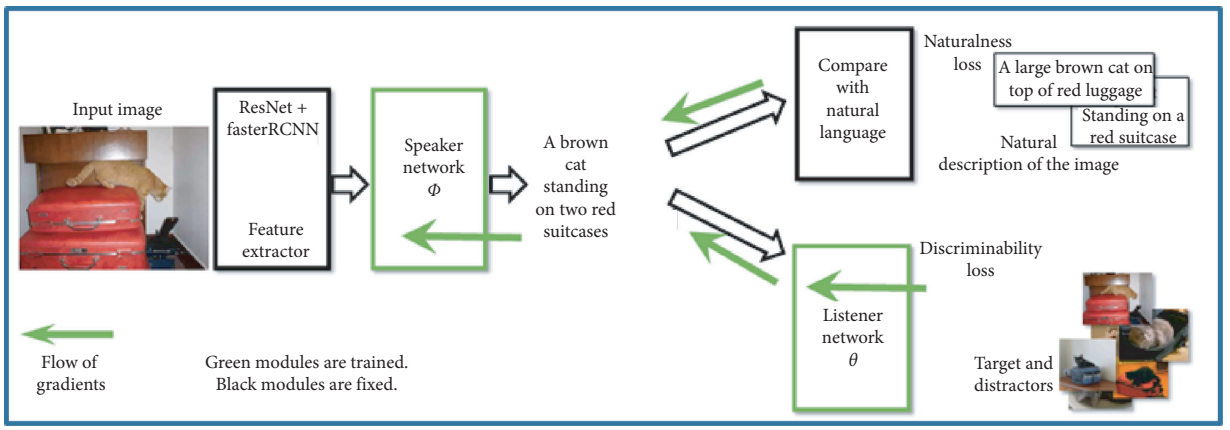

Figure 6: Sample architecture of a multimodal image captioning network [111].

inputs [158]. The limitation of long-distance dependency and inference speed in medical image captioning was tackled using a hierarchical transformer. The model includes an image encoder that extracts features with the support of a bottom-up attention mechanism to capture and extract top-down visual features, as well as a nonrecurrent transformer captioning decoder which helps to compile the generated medical illustration [159]. Salient regions in an image are also extracted using a convolutional model as region features were represented as pooled feature vector. These intraimage region vectors are appropriately attended to obtain suitable weights describing their influence before they are fed into a recurrent model that learns their semantic correlation. The corresponding sequence of the correlated features is transformed into representations which are illustrated as sentences describing the features' interactions [160].

3.1.4. Unsupervised or Semisupervised Captioning. Supervised captioning has so far been productive and successful partly due to the availability of sophisticated deep learning algorithms and an increasing outpour of data. Through the supervised deep learning techniques, a combination of models and frameworks which learn the joint distribution of images and labels has displayed a very intuitive and meaningful illustration of images even similar to humans. However, despite the achievement, the process of completely creating a captioning training set is quite daunting, and the manual effort required to annotate the myriad of images is very challenging. As a result, other means which are free of excessive training data are explored. An unsupervised captioning approach that combines two steps of query and retrieval was researched in [161]. First, several target images are obtained from the internet as well

as a huge database of words describing such images. For any chosen image, words representing its visual display are used to query captions from a reference dataset of sentences. This strategy helps to eliminate manual annotation and also uses multimodal textual-visual information to reduce the effect of noisy words in the vocabulary dataset.

Transfer learning which has seen increasing application in other deep learning domains, especially in computer vision, was applied to image captioning. First, the model is trained on a standard dataset in a supervised manner, and then the knowledge from the supervised model is transferred and applied on a different dataset whose sentences and images are not paired. For this purpose, two autoencoders were designed to train on the textual and visual dataset, respectively, using the distribution of the learned supervised embedding space to infer the unstructured dataset [162]. Also, the process of manual annotation of the training set was semiautomated by evaluating an image into several feature spaces which are individually estimated by an unsupervised clustering algorithm. The centers of the clustered groups are then manually labeled and compiled into a sentence through a voting scheme which compiles all the opinions suggested from each cluster [163]. A set of naïve Bayes model with AdaBoost was used for automatic image annotation by first using a Bayesian classifier to identify unlabeled images and then labeled by a succeeding classifier based on the confidence measurement of the prior classifier [164]. A combination of keywords which have been associated with both labeled and unlabeled images was trained using a graph model. The semantic consistency of the unlabeled images is computed and compared to the labeled images. This is continued until all the unlabeled images are successfully annotated [165].

3.1.5. Difference Captioning. As presented in Figure 7, a spot-the-difference task which describes the differences between two similar images using advance deep learning technique was first investigated in [166]. Their model used a latent variable to capture visual salience in an image pair by aligning pixels which differ in both images. Their work included different model designs such as nearest neighbor matching scheme, captioning masked model, and Difference Description with Latent Alignment uniform for obtaining difference captioning. The Difference Description with Latent Alignment (DDLA) compares both input images at a pixel level via a masked L2 distance function.

Furthermore, the Siamese Difference Captioning Model (SDCM) also combined techniques from deep Siamese convolutional neural network, soft attention mechanism, word embedding, and bidirectional long short-term memory [167]. The features in each image input are computed using the Siamese network, and their differences are obtained using a weighted L1 distance function. Different features are then recursively translated into text using a recurrent neural network and an attention network which focuses on the relevant region on the images. The idea of the Siamese Difference Captioning Model was extended by converting the Siamese encoder into a Fully Convolutional CaptionNet (FCC) through a fully convolutional network [168]. This helps to transform the extracted features into a larger dimension of the input images which makes difference computation more efficient. Also, a word embedding pretrained model was used to embed semantics into the text dataset and beam search technique to ensure multiple options for robustness.

3.2. Datasets. There are several publicly available datasets which are useful for training image captioning tasks. The most popular datasets include Flickr8k [169], Flickr30k [170], MS COCO dataset [83], Visual Genome dataset [171], Instagram dataset [172], and MIT-Adobe FiveK dataset [173].

Flickr30K dataset: it has about 30,000 images from Flickr and about 158,000 captions describing the content of the images. Because of the huge volume of the data, users are able to determine their preferred split size for using the data.

Flickr8K dataset: it has a total of 8,000 images which are divided as 6,000, 1,000, and 1,000 for the training, test, and validation set, respectively. All the images have 5 label captions which are used as a supervised setting for training the images.

Microsoft COCO dataset: it is perhaps the largest captioning dataset, and it also includes training data for object recognition and image segmentation tasks, respectively. The dataset contains around 300,000 images with 5 captions for each image.

3.3. Evaluation Metrics. The automatically generated captions are evaluated to confirm their correctness in describing the given image. In machine learning, some of the common image captioning evaluation measures are as follows.

BLEU (BiLingual Evaluation Understudy) [174]: as a metric, it counts the number of matching n-grams in the model's prediction compared to the ground truth. With this, precision is calculated based on the mean n-grams computed, and the recall is computed via the introduction of a brevity penalty in the caption label.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [175]: it is useful for summary evaluation and is calculated as the overlap of either 1-gram or bigrams between the referenced caption and the predicted sequence. Using the longest sequence available, the co-occurrence F-score mean of the predicted sequence's recall and prediction is obtained.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [176]: it addresses the drawback of BLEU, and it is based on a weighted F-score computation as well as a penalty function meant to check the order of the candidate sequence. It adopts synonyms matching in the detection of similarity between sentences.

CIDEr (Consensus-based Image Description Evaluation) [177]: it determines the consensus between a reference sequence and a predicted sequence via cosine similarity, stemming, and TF-IDF weighting. The predicted sequence is compared to the combination of all available reference sequences.

(a) (b)

Figure 7: Image pair difference annotations of the Spot-the-Diff dataset [166]. (a) The blue truck is no longer there. (b) A car is approaching the parking lot from the right.

Table 2: Class pixel label distribution in the CamVid dataset.

Dataset	Method	B-1	B-2	B-3	B-4	M	C
MS COCO	LSTM-A-2 [179]	0.734	0.567	0.430	0.326	0.254	1.00
	Att-Reg [180]	0.740	0.560	0.420	0.310	0.260	—
	Attend-tell [156]	0.707	0.492	0.344	0.243	0.239	—
	SGC [181]	67.1	48.8	34.3	23.9	21.8	73.3
	phi-LSTM [182]	66.6	48.9	35.5	25.8	23.1	82.1
	COMIC [183]	70.6	53.4	39.5	29.2	23.7	88.1
	TBVA [184]	69.5	52.1	38.6	28.7	24.1	91.9
	SCN [185]	0.741	0.578	0.444	0.341	0.261	1.041
	CLGRU [186]	0.720	0.550	0.410	0.300	0.240	0.960
	A-Penalty [187]	72.1	55.1	41.5	31.4	24.7	95.6
	VD-SAN [188]	73.4	56.6	42.8	32.2	25.4	99.9
	ATT-CNN [189]	73.9	57.1	43.3	33	26	101.6
	RTAN [190]	73.5	56.9	43.3	32.9	25.4	103.3
	Adaptive [191]	0.742	0.580	0.439	0.332	0.266	1.085
	Full-SL [192]	0.713	0.539	0.403	0.304	0.251	0.937
Flickr30K	hLSTMat [193]	73.8	55.1	40.3	29.4	23	66.6
	SGC [181]	61.5	42.1	28.6	19.3	18.2	39.9
	RA + SF [194]	0.649	0.462	0.324	0.224	0.194	0.472
	gLSTM [195]	0.646	0.446	0.305	0.206	0.179	—
	Multi-Mod [196]	0.600	0.380	0.254	0.171	0.169	—
	TBVA [184]	66.6	48.4	34.6	24.7	20.2	52.4
	Attend-tell [156]	0.669	0.439	0.296	0.199	0.185	—
	ATT-FCN [158]	0.647	0.460	0.324	0.230	0.189	—
	VQA [197]	0.730	0.550	0.400	0.280	—	—
	Align-Mod [144]	0.573	0.369	0.240	0.157	—	—
	m-RNN [198]	0.600	0.410	0.280	0.190	—	—
	LRCN [112]	0.587	0.391	0.251	0.165	—	—
	NIC [141]	0.670	0.450	0.300	—	—	—
	RTAN [190]	67.1	48.7	34.9	23.9	20.1	53.3
	3-gated [199]	69.4	45.7	33.2	22.6	23	—
	VD-SAN [188]	65.2	47.1	33.6	23.9	19.9	—
	ATT-CNN [189]	66.1	47.2	33.4	23.2	19.4	—

SPICE (Semantic Propositional Image Caption Evaluation) [178]: it is a relatively new caption metric which relates with the semantic interrelationship between the generated and referenced sequence. Its graph-based methodology uses a scene graph of semantic representations to indicate details of objects and their interaction to describe their textual illustrations.

3.4. Discussion. With an increase in the generation of data, production of sophisticated computing hardware, and complex machine learning algorithms, a lot of achievements have been accomplished in the field of image captioning. Though there have been several implementations, the best results in almost all of the metrics have been recorded through the use of deep learning models. In most cases, the common implementation has been the encoder-decoder architecture which has a feature extractor as the encoder and a language model as the decoder.

Compared to other approaches, this has proven useful as it has become the backbone for more recent designs. To achieve better feature computation, attention mechanism

concepts have been applied to help in focusing on the salient section of images and their features, thereby improving feature-text capturing and translation. In the same manner, other approaches such as generative adversarial network and autoencoders have been thriving in achieving concise image annotation, and to this end, such idea has been incorporated with other unsupervised concepts for captioning purposes as well. For example, reinforced learning technique also generated sequences which are able to succinctly describe images in a timely manner. Furthermore, analyses of several model designs and their results are displayed in Table 2, depicting their efficiency and effectiveness in the BLEU, METEOR, ROUGE-L, CIDEr, and SPICE metrics.

4. Conclusion

In this survey, the state-of-the-art advances in semantic segmentation and image captioning have been discussed. The characteristics and effectiveness of the important techniques have been considered, as well as their process of achieving both tasks. Some of the methods which have accomplished outstanding results have been illustrated including the extraction, identification, and localization of objects in semantic segmentation. Also, the process of feature extraction and transformation into a language model has been studied in the image captioning section. In our estimation, we believe that because of the daunting task of manually segmenting images into semantic classes, as well as the human annotation of images involved in segmentation and captioning, future research would move in the direction of an unsupervised setting of accomplishing this task. This would ensure more energy, and focus is invested solely in the development of complex machine learning algorithms and mathematical models which could improve the present state of the art.

Data Availability

The dataset used for the evalution of the models presented in this study have been discussed in the manuscript as well as their respective references and publications, such as PASCAL VOC: PASCAL Visual Object Classes (VOC) [82], MSCOCO : Microsoft Common Objects in Context [83], Cityscapes : Cityscapes dataset [84], ADE20K : ADE20K dataset [85], CamVid [86], Flickr8k [168], Flickr30k [169], MS COCO Dataset [83], VisualGenome Dataset [170], Instagram Dataset [171], and MIT–Adobe FiveK dataset [172].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the NSFC-Guangdong Joint Fund (Grant no. U1401257), the National Natural Science Foundation of China (Grant nos. 61300090, 61133016, and 61272527), the Science and Technology Plan Projects in Sichuan Province (Grant no. 2014JY0172), and the Opening Project of Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology (Grant no. 2013A061401003).

References

[1] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. M. Farinella, "Computer vision for assistive technologies," *Computer Vision and Image Understanding*, vol. 154, pp. 1–15, 2017.

[2] A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman, "Crowdsourcing in computer vision," *Foundations and Trends in Computer Graphics and Vision*, vol. 10, no. 3, pp. 177–243, 2016.

[3] H. Ayaz, M. Ahmad, M. Mazzara, and A. Sohaib, "Hyperspectral imaging for minced meat classification using nonlinear deep features," *Applied Sciences*, vol. 10, no. 21, p. 7783, 2020.

[4] A. Oluwasanmi, M. U. Aftab, A. Shokanbi, J. Jackson, B. Kumeda, and Z. Qin, "Attentively conditioned generative adversarial network for semantic segmentation," *IEEE Access*, vol. 8, pp. 31733–31741, 2020.

[5] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Proceedings of the 2017 Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.

[7] B. G. Weinstein, "A computer vision for animal ecology," *Journal of Animal Ecology*, vol. 87, no. 3, pp. 533–545, 2018.

[8] A. Oluwasanmi, M. U. Aftab, Z. Qin et al., "Transfer learning and semisupervised adversarial detection and classification of COVID-19 in CT images," *Complexity*, vol. 2021, Article ID 6680455, 11 pages, 2021.

[9] M. Ahmad, I. Haq, Q. Mushtaq, and M. Sohaib, "A new statistical approach for band clustering and band selection using k-means clustering," *International Journal of Engineering and Technology*, vol. 3, 2011.

[10] M. Buckler, S. Jayasuriya, and A. Sampson, "Reconfiguring the imaging pipeline for computer vision," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 975–984, Venice, Italy, October 2017.

[11] M. Leo, A. Furnari, G. G. Medioni, M. M. Trivedi, and G. M. Farinella, "Deep learning for assistive computer vision," in *Proceedings of the European Conference on Computer Vision Workshops*, Munich, Germany, September 2018.

[12] H. Fang, S. Gupta, F. N. Iandola et al., "From captions to visual concepts and back," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1473–1482, Boston, MA, USA, June 2015.

[13] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine Learning*, vol. 3, no. 2/3, pp. 95–99, 1988.

[14] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge, UK, 2014.

[15] E. Alpaydin, "Introduction to machine learning," *Adaptive Computation And Machine Learning*, MIT Press, Cambridge, MA, USA, 2004.

[16] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in

Proceedings of the Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, August 2012.

[17] I. G. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning. Nature*, vol. 521, pp. 436–444, 2015.

[18] F. Hutter, L. Kotthoff, and J. Vanschoren, "Automated machine learning.: methods, systems, challenges," *Automated Machine Learning*, MIT Press, Cambridge, MA, USA, 2019.

[19] R. Singh, A. Sonawane, and R. Srivastava, "Recent evolution of modern datasets for human activity recognition: a deep survey," *Multimedia Systems*, vol. 26, 2020.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the 2011 International Conference on Computer Vision*, pp. 2556–2563, Barcelona, Spain, November 2011.

[21] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput*, vol. 28, pp. 976–990, 2011.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proceedings of the IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1597–1600, Boston, MA, USA, August 2017.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556.

[25] O. Ivanov, M. Figurnov, and D. P. Vetrov, "Variational autoencoder with arbitrary conditioning," in *Proceedings of the International Conference on Learning Representations*, Vancouver, Canada, May 2018.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[28] K. Khan, M. S. Siddique, M. Ahmad, and M. Mazzara, "A hybrid unsupervised approach for retinal vessel segmentation," *BioMed Research International*, vol. 2020, Article ID 8365783, 20 pages, 2020.

[29] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.

[30] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, Honolulu, HI, USA, July 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[32] H. Hassan, A. Bashir, M. Ahmad et al., "Real-time image dehazing by superpixels segmentation and guidance filter," *Journal of Real-Time Image Processing*, 2020.

[33] C. Liu, L. Chen, F. Schroff et al., "Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

[34] A. Kirillov, K. He, R. B. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

[35] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Medical Image Analysis*, vol. 54, pp. 88–99, 2019.

[36] A. Arnab, S. Zheng, S. Jayasumana et al., "Conditional random fields meet deep neural networks for semantic segmentation: combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 37–52, 2018.

[37] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, pp. 146–168, 2004.

[38] A. Oluwasanmi, Z. Qin, and T. Lan, "Brain MR segmentation using a fusion of K-means and spatial Fuzzy C-means," in *Proceeding of International Conference on Computer Science and Application Engineering*, Wuhan, China, July 2017.

[39] A. Oluwasanmi, Z. Qin, T. Lan, and Y. Ding, "Brain tissue segmentation in MR images with FGM," in *Proceeding of the International Conference on Artificial Intelligence and Computer Science*, Guilin, China, December 2016.

[40] J. Chen, C. Yang, G. Xu, and L. Ning, "Image segmentation method using Fuzzy C mean clustering based on multi-objective optimization," *Journal of Physics: Conference Series*, vol. 1004, pp. 12–35, 2018.

[41] A. Oluwasanmi, Z. Qin, and T. Lan, "Fusion of Gaussian mixture model and spatial Fuzzy C-means for brain MR image segmentation," in *Proceedings of International Conference on Computer Science and Application Engineering*, Wuhan, China, July 2017.

[42] B. J. Liu and L. Cao, "Superpixel segmentation using Gaussian mixture model," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4105–4117, 2018.

[43] B. Kang and T. Q. Nguyen, "Random forest with learned representations for semantic segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3542–3555, 2019.

[44] Z. Zhao and X. Wang, "Multi-egments Naïve Bayes classifier in likelihood space," *IET Computer Vision*, vol. 12, no. 6, pp. 882–891, 2018.

[45] T. Y. Tan, L. Zhang, C. P. Lim, B. Fielding, Y. Yu, and E. Anderson, "Evolving ensemble models for image segmentation using enhanced particle swarm optimization," *IEEE Access*, vol. 7, pp. 34004–34019, 2019.

[46] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2354–2367, 2018.

[47] S. Mohapatra, "Segmentation using support vector machines," in *Proceedings of the Second International Conference on Advanced Computational and Communication Paradigms (ICACCP 2019)*, pp. 1–4, Gangtok, India, November 2019.

[48] Ç. Kaymak and A. Uçar, "A brief survey and an application of semantic image segmentation for autonomous driving," *Handbook of Deep Learning Applications*, Springer, Berlin, Germany, 2018.

[49] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, September 2014.

[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[51] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4438–4446, Honolulu, HI, USA, July 2017.

[52] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training," in *Proceedings of the European Conference on Computer Vision*, pp. 381–397, Amsterdam, The Netherlands, October 2016.

[53] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[54] N. Wang, S. Li, A. Gupta, and D. Yeung, "Transferring rich feature hierarchies for robust visual tracking," 2015, http://arxiv.org/abs/1501.04587.

[55] R. B. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[56] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[57] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Venice, Italy, October 2017.

[58] A. Salvador, X. Giró, F. Marqués, and S. Satoh, "Faster R-CNN features for instance search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2016)*, pp. 394–401, Las Vegas, NV, USA, June 2016.

[59] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Boston, MA, USA, June 2015.

[60] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[61] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, Santiago, Chile, December 2015.

[62] O. Ronneberger, P. Fischer, T. Brox, and U-Net, *Convolutional Networks for Biomedical Image Segmentation*, MICCAI, Munich, Germany, 2015.

[63] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. J. Pal, "The importance of skip connections in biomedical image segmentation," 2016, http://arxiv.org/abs/1608.04117.

[64] S. Jégou, M. Drozdzal, D. Vázquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, July 2017.

[65] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: gated shape CNNs for semantic segmentation," 2019, http://arxiv.org/abs/1907.05740.

[66] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, http://arxiv.org/abs/1511.07122.

[67] P. Bilinski and V. Prisacariu, "Dense decoder shortcut connections for single-pass semantic segmentation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6596–6605, Salt Lake City, UT, USA, June 2018.

[68] Z. Zhang, X. Zhang, C. Peng, D. Cheng, and J. Sun, "ExFuse: enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, September 2018.

[69] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision*, Kolding, Denmark, June 2017.

[70] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: deep feature aggregation for real-time semantic segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

[71] W. Xiang, H. Mao, and V. Athitsos, "ThunderNet: a turbo unified network for real-time semantic segmentation," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1789–1796, Village, HI, USA, January 2019.

[72] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proceedings of the International Conference on Learning Representations*, pp. 11–25, San Diego, CA, USA, May 2015.

[73] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2016.

[74] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, http://arxiv.org/abs/1706.05587.

[75] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: rethinking dilated convolution in the backbone for semantic segmentation," 2019, http://arxiv.org/abs/1903.11816.

[76] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, Salt Lake City, UT, USA, June 2018.

[77] P. H. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with Convolutional Networks," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1713–1721, Boston, MA, USA, June 2015.

[78] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele, "Simple does it: weakly supervised instance and semantic segmentation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1665–1674, Honolulu, HI, USA, July 2017.

[79] G. Papandreou, L. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1742–1750, Santiago, Chile, December 2015.

[80] J. Dai, K. He, and J. Sun, "BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1635–1643, Santiago, Chile, December 2015.

[81] W. Hung, Y. Tsai, Y. Liou, Y. Lin, and M. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proceedings of the British Machine Vision Conference*, Newcastle, UK, September 2018.

[82] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2009.

[83] T. Lin, M. Maire, S. J. Belongie et al., "Microsoft COCO: common objects in context," *ECCV*, pp. 740–755, Springer, Berlin, Germany, 2014.

[84] M. Cordts, M. Omran, S. Ramos et al., "The Cityscapes dataset," in *Proceedings of the CVPR Workshop on the Future of Datasets in Vision*, Boston, MA, USA, June 2015.

[85] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, Honolulu, HI, USA, July 2017.

[86] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: the KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[87] G. Ros, L. Sellart, J. Materzynska, D. Vázquez, and A. M. López, "The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3234–3243, Las Vegas, NV, USA, June 2016.

[88] S. Nowozin, "Optimal decisions from probabilistic models: the intersection-over-union case," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 548–555, Columbus, OH, USA, June 2014.

[89] C. Wu, H. Cheng, S. Li, H. Li, and Y. Chen, "ApesNet: a pixel-wise efficient segmentation network for embedded devices," in *Proceedings of the 14th ACM/IEEE Symposium on Embedded Systems for Real-Time Multimedia*, pp. 1–7, Pittsburgh, PA, USA, October 2016.

[90] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: a deep neural network architecture for real-time semantic segmentation," 2016, http://arxiv.org/abs/1606.02147.

[91] A. Chaurasia and E. Culurciello, "LinkNet: exploiting encoder representations for efficient semantic segmentation," in *Proceedings of the IEEE Visual Communications and Image Processing*, pp. 1–4, St. Petersburg, FL, USA, December 2017.

[92] L. Fan, W. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, 2018.

[93] C. Yu, J. Wang, G. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, September 2018.

[94] J. Li, Y. Zhao, J. Fu, J. Wu, and J. Liu, "Attention-guided network for semantic video segmentation," *IEEE Access*, vol. 7, pp. 140680–140689, 2019.

[95] H. Zhou, K. Song, X. Zhang, W. Gui, and Q. Qian, "WAILS: watershed algorithm with image-level supervision for weakly supervised semantic segmentation," *IEEE Access*, vol. 7, pp. 42745–42756, 2019.

[96] L. Zhang, P. Shen, G. Zhu et al., "Improving semantic image segmentation with a probabilistic superpixel-based dense conditional random field," *IEEE Access*, vol. 6, pp. 15297–15310, 2018.

[97] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3376–3385, Boston, MA, USA, June 2015.

[98] C. Han, Y. Duan, X. Tao, and J. Lu, "Dense convolutional networks for semantic segmentation," *IEEE Access*, vol. 7, pp. 43369–43382, 2019.

[99] R. Vemulapalli, O. Tuzel, M. Y. Liu, and R. Chellapa, "Gaussian conditional random field network for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3233, Las Vegas, NV, USA, June 2016.

[100] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1377–1385, Santiago, Chile, December 2015.

[101] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3194–3203, Las Vegas, NV, USA, July 2016.

[102] C. X. Peng, X. Zhang, K. Jia, G. Yu, and J. Sun, "MegDet: a large mini-batch object detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6181–6189, Salt Lake City, UT, USA, June 2008.

[103] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, pp. 519–534, Amsterdam, The Netherlands, October 2016.

[104] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4151–4160, Honolulu, HI, USA, July 2017.

[105] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5168–5177, Honolulu, HI, USA, July 2017.

[106] H.-H. Han and L. Fan, "A new semantic segmentation model for supplementing more spatial information," *IEEE Access*, vol. 7, pp. 86979–86988, 2019.

[107] X. Jin, X. Li, H. Xiao et al., "Video scene parsing with predictive feature learning," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5581–5589, Venice, Italy, October 2017.

[108] P. Wang, P. Chen, Y. Yuan et al., "Understanding convolution for semantic segmentation," in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451–1460, Lake Tahoe, NV, USA, March 2018.

[109] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, Honolulu, HI, USA, June 2017.

[110] G. Vered, G. Oren, Y. Atzmon, and G. Chechik, "Cooperative image captioning," 2019, http://arxiv.org/abs/1907.11565.

[111] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, and S. Venugopalan, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, Boston, MA, USA, June 2015.

[112] Z. Fan, Z. Wei, S. Wang, and X. Huang, *Bridging by Word: Image Grounded Vocabulary Construction for Visual Captioning*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019.

[113] Y. Zhou, Y. Sun, and V. Honavar, "Improving image captioning by leveraging knowledge graphs," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 283–293, Waikoloa Village, HI, USA, January 2019.

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

[115] Q. Wang and A. B. Chan, "Describing like humans: on diversity in image captioning," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

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

[117] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 10039–10042, Yokohama, Japan, August 2019.

[118] S. Wang, L. Lan, X. Zhang, G. Dong, and Z. Luo, "Cascade semantic fusion for image captioning," *IEEE Access*, vol. 7, pp. 66680–66688, 2019.

[119] Y. Su, Y. Li, N. Xu, and A. Liu, "Hierarchical deep neural network for image captioning," *Neural Processing Letters*, vol. 52, pp. 1–11, 2019.

[120] M. Yang, W. Zhao, W. Xu et al., "Multitask learning for cross-domain image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2019.

[121] Z. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[122] S. Sheng, K. Laenen, and M. Moens, "Can image captioning help passage retrieval in multimodal question answering?" in *Proceedings of European Conference on Information Retrieval (ECIR)*, pp. 94–101, Springer, Cologne, Germany, April 2019.

[123] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2743–2754, 2019.

[124] S. Sreedevi and S. Sebastian, "Content based image retrieval based on Database revision," in *Proceedings of the International Conference on Machine Vision and Image Processing*, pp. 29–32, Taipei, Taiwan, December 2012.

[125] E. R. Vimina and J. K. Poulose, "Image retrieval using colour and texture features of Regions of Interest," in *Proceedings of the International Conference on Information Retrieval and Knowledge Management*, pp. 240–243, Kuala Lumpur, Malaysia, December 2012.

[126] V. Ordonez, G. Kulkarni, and T. R. Berg, "Im2text: describing images using 1 million captioned photographs," in *Advances in Neural Information Processing Systems*, pp. 1143–1151, Springer, Berlin, Germany, 2011.

[127] J. R. Curran, S. Clark, and J. Bos, "Linguistically motivated large-scale NLP with C and C and boxer," in *Proceedings of the Forty Fifth Annual Meeting of the ACL on Inter-Active Poster and Demonstration Sessions*, pp. 33–36, Prague, Czech, June 2007.

[128] D. R. Hardoon, S. R. Szedmak, J. R. Shawe-Taylor, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[129] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[130] J. Wang, Y. Zhao, Q. Qi et al., "MindCamera: interactive sketch-based image retrieval and synthesis," *IEEE Access*, vol. 6, pp. 3765–3773, 2018.

[131] D. Xu, X. Alameda-Pineda, J. Song, E. Ricci, and N. Sebe, "Cross-paced representation learning with partial curricula for sketch-based image retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4410–4421, 2018.

[132] K. Song, F. Li, F. Long, J. Wang, J. Wang, and Q. Ling, "Discriminative deep feature learning for semantic-based image retrieval," *IEEE Access*, vol. 6, pp. 44268–44280, 2018.

[133] P. Kuznetsova, V. Ordonez, A. C. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions," *Association for Computational Linguistics*, vol. 1, pp. 359–368, 2012.

[134] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TREETALK: composition and compression of trees for image descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, no. 10, pp. 351–362, 2014.

[135] M. Mitchell, "Midge: generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756, Avignon, France, April 2012.

[136] Y. Yang, C. L. Teo, H. Daumé, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 444–454, Portland, OR, USA, June 2011.

[137] A. Kojima, M. Izumi, T. Tamura, and K. Fukunaga, "Generating natural language description of human behavior from video images," in *Proceedings of the ICPR 2000*, vol. 4, pp. 728–731, Barcelona, Spain, September 2000.

[138] A. Kojima, T. Tamura, and K. Fukunaga, "natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.

[139] A. Tariq and H. Foroosh, "A context-driven extractive framework for generating realistic image descriptions," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 619–632, 2017.

[140] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," 2014, http://arxiv.org/abs/1411.4555.

[141] J. M. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: fully convolutional localization networks for dense captioning," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4565–4574, Las Vegas, NV, US, June 2016.

[142] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 271–275, New York, NY, USA, June 2016.

[143] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," 2014, http://arxiv.org/abs/1412.2306.

[144] R. Krishna, K. Hata, F. Ren, F. Li, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, Venice, Italy, October 2017.

[145] L. Yang, K. D. Tang, J. Yang, and L. Li, "Dense captioning with joint inference and visual context," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1978–1987, Honolulu, HI, USA, July 2017.

[146] G. Srivastava and R. Srivastava, *A Survey on Automatic Image Captioning. International Conference on Mathematics and Computing*, Springer, Berlin, Germany, 2018.

[147] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, "Image captioning with both object and scene information," in *ACM Multimedia,* Springer, Berlin, Germany, 2016.

[148] W. Jiang, L. Ma, Y. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proceedings of the ECCV*, Munich, Germany, September 2018.

[149] Q. Wang and A. B. Chan, "CNN+CNN: convolutional decoders for image captioning," 2018, http://arxiv.org/abs/1805.09019.

[150] K. Zheng, C. Zhu, S. Lu, and Y. Liu, *Multiple-Level Feature-Based Network for Image Captioning*, Springer, Berlin, Germany, 2018.

[151] X. Li and Q. Jin, *Improving Image Captioning by Concept-Based Sentence Reranking*, Springer, Berlin, Germany, 2016.

[152] T. Karayil, A. Irfan, F. Raue, J. Hees, and A. Dengel, *Conditional GANs for Image Captioning with Sentiments*, ICANN, Los Angeles, CA, USA, 2019.

[153] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the NIPS*, Long Beach, CA, USA, December 2017.

[154] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, January 2016.

[155] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, Lille, France, July 2015.

[156] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015, http://arxiv.org/abs/1409.0473.

[157] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4651–4659, Las Vegas, NV, USA, June 2016.

[158] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Proceedings of the MLMI@ MICCAI*, Shenzhen, China, October 2019.

[159] S. Wang, H. Mo, Y. Xu, W. Wu, and Z. Zhou, "Intra-image region context for image captioning," *PCM*, Springer, Berlin, Germany, 2018.

[160] L. Pellegrin, J. A. Vanegas, J. Arevalo et al., "A two-step retrieval method for image captioning," *Lecture Notes in Computer Science*, pp. 150–161, Springer, Berlin, Germany, 2016.

[161] A. Carraggi, M. Cornia, L. Baraldi, and R. Cucchiara, "Visual-semantic alignment across domains using a semi-supervised approach," in *Proceedings of the European Conference on Computer Vision Workshops*, pp. 625–640, Munich, Germany, September 2018.

[162] S. Vajda, D. You, S. Antani, and G. Thoma, "Large image modality labeling initiative using semi-supervised and optimized clustering," *International Journal of Multimedia Information Retrieval*, vol. 4, no. 2, pp. 143–151, 2015.

[163] H. M. Castro, L. E. Sucar, and E. F. Morales, "Automatic image annotation using a semi-supervised ensemble of classifiers," *Lecture Notes in Computer Science*, vol. 4756, pp. 487–495, Springer, Berlin, Germany, 2007.

[164] Y. Xiao, Z. Zhu, N. Liu, and Y. Zhao, "An interactive semi-supervised approach for automatic image annotation," in *Proceedings of the Pacific-Rim Conference on Multimedia*, pp. 748–758, Singapore, December 2012.

[165] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, October 2018.

[166] A. Oluwasanmi, M. U. Aftab, E. Alabdulkreem, B. Kumeda, E. Y. Baagyere, and Z. Qin, "CaptionNet: automatic end-to-end siamese difference captioning model with attention," *IEEE Access*, vol. 7, pp. 106773–106783, 2019.

[167] A. Oluwasanmi, E. Frimpong, M. U. Aftab, E. Y. Baagyere, Z. Qin, and K. Ullah, "Fully convolutional CaptionNet: siamese difference captioning attention model," *IEEE Access*, vol. 7, pp. 175929–175939, 2019.

[168] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[169] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, Santiago, Chile, December 2015.

[170] R. Krishna, Y. Zhu, O. Groth et al., "Visual genome: connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[171] K. Tran, X. He, L. Zhang, and J. Sun, "Rich image captioning in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 434–441, Las Vegas, NA, USA, July 2016.

[172] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," *Computer Vision and Pattern Recognition (CVPR)*, vol. 97, 2011.

[173] K. Papineni, S. Roukos, T. Ward, and W. Zhu, *Bleu: A Method For Automatic Evaluation Of Machine Translation*, pp. 311–318, Association for Computational Linguistics, Stroudsburg, PA, USA, 2001.

[174] C. Lin, *ROUGE: A Package For Automatic Evaluation Of Summaries*, pp. 74–81, Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 2004.

[175] S. Banerjee and A. Lavie, "METEOR: an automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings Of The Meeting Of The Association For Computational Linguistics*, pp. 65–72, Ann Arbor, MI, USA, June 2005.

[176] R. Vedantam, C. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in *Proceedings Of The Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, Boston, MA, USA, June 2015.

[177] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: semantic propositional image caption evaluation," in *Proceedings Of The European Conference on Computer Vision*, pp. 382–398, Springer, Amsterdam, The Netherlands, October 2016.

[178] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings Of The IEEE International Conference on Computer Vision (ICCV)*, pp. 4904–4912, Venice, Italy, October 2017.

[179] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1367–1381, 2018.

[180] N. Xu, A.-A. Liu, J. Liu et al., "Scene graph captioner: image captioning based on structural visual representation," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 477–485, 2019.

[181] Y. H. Tan, C. S. Chan, and C. S. Chan, "Phrase-based image caption generator with hierarchical LSTM network," *Neurocomputing*, vol. 333, pp. 86–100, 2019.

[182] J. H. Tan, C. S. Chan, and J. H. Chuah, "COMIC: towards a compact image captioning model with attention," *IEEE Transactions on Multimedia*, vol. 99, 2019.

[183] C. He and H. Hu, "Image captioning with text-based visual attention," *Neural Processing Letters*, vol. 49, no. 1, pp. 177–185, 2019.

[184] Z. Gan, C. Gan, X. He et al., "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, June 2017.

[185] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language cnn for image captioning," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[186] J. Li, M. K. Ebrahimpour, A. Moghtaderi, and Y.-Y. Yu, "Image captioning with weakly-supervised attention penalty," 2019, http://arxiv.org/abs/1903.02507.

[187] X. He, B. Yang, X. Bai, and X. Bai, "VD-SAN: visual-densely semantic attention network for image caption generation," *Neurocomputing*, vol. 328, pp. 48–55, 2019.

[188] D. Zhao, Z. Chang, S. Guo, Z. Chang, and S. Guo, "A multimodal fusion approach for image captioning," *Neurocomputing*, vol. 329, pp. 476–485, 2019.

[189] W. Wang and H. Hu, "Image captioning using region-based attention joint with time-varying attention," *Neural Processing Letters*, vol. 13, 2019.

[190] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via A visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[191] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[192] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, 2019.

[193] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 2016.

[194] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE in- Ternational Conference on Computer Vision*, pp. 2407–2415, Santiago, Chile, December 2015.

[195] R. Kiros, R. Zemel, and R. Salakhutdinov, "Multimodal neural language models," in *Proceedings of the International Conference on Machine Learning*, Beijing, China, June 2014.

[196] Q. Wu, C. Shen, P. Wang, A. Dick, and A. V. Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1367–1381, 2016.

[197] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks," in *Proceedings of the International Conference on Learning Representation*, San Diego, CA, USA, May 2015.

[198] A. Yuan, X. Li, and X. Lu, "3G structure for image caption generation," *Neurocomputing*, vol. 330, pp. 17–28, 2019.

[199] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proceedings of the ECCV*, pp. 44–57, Marseille, France, October 2008.