Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Survey on Referring Image Segmentation

Honglin Wang *

International College, Zhengzhou University, Zhengzhou, Henan Province, China

* Corresponding Author Email: leowp@stu.zzu.edu.cn

Abstract. With the popularity of artificial intelligence models and the increasing expectation of artificial intelligence applications in many fields, reference image segmentation (RIS) has attracted much attention from researchers. RIS, as one of the most basic and challenging visual language cross-modal tasks in the intersection of computer vision and natural language processing, aims to segment an instance from an image corresponding to a given natural language representation. This paper aims to provide an overview as comprehensive as possible, covering the mainstream benchmark datasets and their statistic information, common evaluation metrics, a few crucial and representative works in RIS, and the performance evaluation of each proposed method. Included RIS methods are elaborated with their core model structure and procedure in performing RIS, and are categorized into 5 classes in this paper based on how multimodal information is processed. At the end of this paper, the author makes a brief expectation of possible future expansions on the research of RIS.

Keywords: Referring image segmentation; deep learning; computer vision; natural language processing.

1. Introduction

In recent years, with the popularization and intelligentization of camera equipment, all kinds of image collecting equipment can be widely applied, the cost of image acquisition is greatly reduced, and a large amount of images and video data are generated every moment. With the development of Internet technology, images, as a carrier of rich visual information, are widely disseminated on the Internet through people's sharing. Meanwhile iterations of high-performance image processing equipment have driven the application of techniques such as image classification, object detection, and image segmentation.

In order to accurately represent goals and requirements to guide computers to accomplish tasks involving localization, people introduce computers with referential expression (RE) and related tasks. The task of referring image segmentation (RIS) is: given an input image and an RE, the computer analyzes and extracts features from both inputs, then outputs the region of instance described by RE, which is called an image segment or mask. The referring expression here is the accurate linguistic representation of the unique features of a particular described instance or region in the image, such as "partial guy in background facing us", as shown in Fig 1.

Fig. 1 The task of RIS is to predict a mask of the instance based on image and referring expression. Each instance could be referred by a few different expressions.

Similar to semantic segmentation and instance segmentation, RIS needs computers to determine the instance location, then subsequently refine the instance outline to a pixel-level precision. However, they differ in that segmentation is done on a specified instance by RE in RIS, instead of being performed on all regions and instances of any category.

RIS can be applied to related fields where precision guide or localization work is required. For instance, in medical image analysis, it may assist in tumor detection, organ image segmentation, blood vessel analysis, etc.; in autonomous driving and intelligent transportation systems, intelligent cars may collect the images of road and pedestrian and perform automatic route planning; in visual effects and image editing, RIS can assist in background extraction, object replacement, and image synthesis, etc. [1,2]; in security systems, it can be used for video surveillance such as target tracking [3], etc.; in smart home, it can enable intelligent robots to correctly complete tasks that require localization, such as navigation, handling, etc.; in entertainment and life, it can be used in human-computer interaction [1], virtual reality, augmented reality, and other technologies.

For computers, RIS is a complicated and difficult task to handle. Firstly, for texts, computers need to accurately understand the referring expressions in order to guide RIS. However, the length of the RE is not fixed, it may have convoluted and confusing grammatical structures and implicit contextual correlations, making it difficult to parse the texts. Secondly, for images, computer-acquired images may have some problems, such as significant noise, complex scenes and distortion, bad lighting conditions, object overlap and occlusion, blurred boundaries, etc., making it difficult for computers to capture structured information. Thirdly, during the training process of the model, it is necessary for the computer to obtain data with different modal information. Data annotation is very time-consuming and requires professional operation, which does not allow the introduction of any possible errors and requires that the number of different classes of instances in the data tends to be balanced, hence obtaining high-quality and sufficient quantity of data could be a challenge. Another challenge is to enable computers to integrate and match information from text and images, e.g., such as whether the model is trained to be capable of detecting and returning an empty segmentation when the image does not contain instances corresponding to the referring expressions. Last but not least, RIS will face the challenge of real-time application. For example, in the autopilot system, the algorithm needs to maintain high accuracy, high robustness and high efficiency to respond quickly and appropriately to the sudden changes in the signal lights and the surrounding environment.

2. RIS Datasets and Evaluation Metrics

2.1. RIS Datasets

The dataset used in the training of RIS tasks needs to include images, pixel-level annotations of one or more instances in the image, and referring expressions corresponding to the instances. The commonly used image RIS datasets in recent years are shown in Table 1, where the training (train), validation (val), and test sets columns refer to the cardinal of the referring expression set partitions. These datasets are derived from different image retrieval databases.

2.1.1. ReferIt

The research on Referring Expression Generation can be traced back to the 1970s [4]. As the research on referring expression and multimodal learning grows, researchers need datasets targeted at RIS. In 2006, Grubinger et al. collected Image CLEF IAPR including the IAPR TC-12 [5]. IAPR TC-12 [5] is a retrieval database containing 20,000 images without copyright restrictions. Subsequently later in 2010, Escalante et al. used ISATOOL to create SAIAPR TC-12 [6], which introduced segmentation of 99,535 instances, but each instance was only matched to a label instead of an RE.

In 2014, Kazemzadeh Sahar et al. [4] Constructed the first large-scale RIS dataset. The REs of this dataset were collected through a 2-player game called ReferItGame. Player 1 is asked to type in a text box a referring expression for a given segmented instance in the shown image, then Player 2 needs to click on the instance in a segment-free image based on the RE written by Player 1. If Player 2 clicks

correctly, both players will receive game points and exchange their game roles for the next round, otherwise the players will not be awarded and will keep their current roles.

2.1.2. MS COCO

MS COCO (Microsoft Common Objects in Context) is a large-scale image retrieval dataset funded by Microsoft in 2014 [7], which was collected by Tsung-Yi Lin et al. and whose images were annotated through ReferItGame. The datasets presented below are all constructed on MS COCO.

RefCOCO, RefCOCO+ [8]. These 2 common datasets are comprised of images randomly chosen from MS COCO. RefCOCO+ differs from RefCOCO in terms of the RE, by that RefCOCO+ does not allow absolute positional words, such as “left”. By different split allocations, the same dataset may have different training, validation and test sets. UNC (University of North Carolina) divided their test set into two categories based on images, with testA and testB contains only people and only non-people instances respectively. Google uses different partitioning method, its train, val, test sets are divided in terms of instances, not images.

RefCOCOg [9]. A variant of RefCOCO. It was annotated in a non-interactive setting: one group of personnel provides REs with respect to the provided instances, the other group selects the correspondent instance region. This process was repeated 3 times in a cross-validation manner. RefCOCOg requires the image to contain 2 to 4 objects of the same category, with a combined coverage area exceeding 5% of that of the image. The average length of the REs in RefCOCO is approximately 3.5 words, whereas in RefCOCOg it is about 8.4 words. There are 2 common partition of RefCOCOg, constructed seperately by UMD (University of Maryland) and Google, which differs in that Google split has no canonical test set.

GuessWhat [10]. REs in GuessWhat was collected through a 2-player cooperative game. Both players are presented with an image containing multiple objects, Player 1 is assigned a random instance, and Player 2 performs a series of yes/no questions to determine the target object.

gRefCOCO (GeneralizedRefCOCO) [11]. This dataset includes 80,022 multi-instance expressions and 32,202 non-instance expressions, e.g., "Everyone except the kid in white" may refer to multiple targets, or it may refer to zero targets in an image without anyone. Its REs bear more complex structures, e.g. "the bike that has two passengers and its driver" describes a complex pertinence among the instances.

Table 1. Statistics of the datasets for RIS. The partition reference is attached to dataset names.

Dataset	Source	Images	Instances	Expressions	train	val	test
ReferIt	CLEF	19894	96654	130525	59976	10444	60105
RefCOCO(UNC)	MSCOCO	19994	50000	142209	120624	10834	5656/5095
RefCOCO+(UNC)	MSCOCO	19992	49856	141564	120191	10758	5726/4889
RefCOCOg(Google)	MSCOCO	25799	49822	95010	85474	9536	-
Guess What	MSCOCO	66537	134073	821889	-	-	-
gRefCOCO	MSCOCO	19994	60287	278232	-	-	-

2.2. RIS Evaluation Metrics

In order to comparatively evaluate the performance of different RIS models such as accuracy, generalization ability, stability and robustness, researchers can inspect the models on different datasets with appropriate evaluation metrics. The accuracy of segmentation can be reflected by calculating the difference from predicted segmentation to the factual segmentation (i.e. ground truth). Commonly used metrics shown below are Intersection over Union (IoU) (1) [12], Overall IoU (oIoU) (2), g-IoU (3) [13], d-IoU (4) [14], c-IoU (5) [14], Dice coefficent (6), and its generalized version, Tversky coefficent (7) [15]. In general, these values are taken as the average on the dataset, and the higher these values are, the more accurate the model.

$$\mathrm{IoU} = \frac{|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P} \cup \mathrm{G}|} \tag{1}$$

$$\mathrm{oIoU} = \frac{\sum_i |\mathrm{P_i} \cap \mathrm{G_i}|}{\sum_i |\mathrm{P_i} \cup \mathrm{G_i}|} \tag{2}$$

$$\mathrm{gIoU} = \mathrm{IoU} - \frac{|\mathrm{C} - (\mathrm{P} \cup \mathrm{G})|}{|\mathrm{C}|} \tag{3}$$

$$\mathrm{dIoU} = \mathrm{IoU} - \frac{\rho^2}{\mathrm{c}^2} \tag{4}$$

$$\begin{cases} \mathrm{cIoU} = \mathrm{dIoU} - \dfrac{\mathrm{v}^2}{(1 - \mathrm{IoU}) + \mathrm{v}} \\ \mathrm{v} = \dfrac{4}{\pi^2}\left(\tan^{-1}\dfrac{\mathrm{w_P}}{\mathrm{h_p}} - \tan^{-1}\dfrac{\mathrm{w_G}}{\mathrm{h_G}}\right)^2 \end{cases} \tag{5}$$

$$Dice = \frac{2|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P}| + |\mathrm{G}|} \tag{6}$$

$$Tversky = \frac{|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P} \cap \mathrm{G}| + \alpha|\mathrm{P} - \mathrm{G}| + \beta|\mathrm{G} - \mathrm{P}|} \tag{7}$$

Where P and G are the predicted segmentation and ground truth respectively, ρ is their geometric center distance, C is their minimal convex enclosure, c is the maximum diameters (diagonal lengths for rectangles) of C, w_i and h_i are the widths and heights of the segmentation i (maximum length in horizontal and vertical directions), α and β are predefined constants. The Dice coefficient is a special case of Tversky coefficient when$\alpha = \beta = 0.5$, which is equivalent to the metric F1-score.

Another family of commonly used evaluation metric is the metric percentage, metrix@x. A widely-used representative of this family is called the precision percentage Prec@x, which indicates the percentage of instances in the test set whose IoU exceeds a threshold x. For instance, Prec@0.5 is the percentage of all segmentations with IoU greater than 0.5. The commonly used values for x are 0.5, 0.6, 0.7, 0.8, and 0.9. Other metrics of this family include Recall@x, F1-score@x, etc.

3. Overview of RIS Methods

The fundamental problem RIS faced is how to match image and text information. The encoding method and matching mechanism of information are the keys to enable computers to simultaneously comprehend information from both modalities. For visual information, commonly used encoding methods include graph theory methods and CNNs (Convolutional Neural Networks). As for linguistic information, common encoders are the RNN and Transformer classes. The matching mechanism is the further combination processing of encoded or extracted features from both modalities, such like mapping both features to the same vector subspace and adding them successively, so as to capture their correlation and dependency. This section will introduce 5 classes of existing major RIS methods, some of which have intersections in different categories. This article takes their core methods as the main reference in classification.

3.1. Direct Fusion Methods

Multimodal fusion technology is a matching mechanism for models to process multimodal data, which mainly fuses the multimodal information through methods e.g. Concatenate, Joint, Coordinate, and Encode-Decode [16], and then applies fused features to subsequent tasks. The method based on the CNN-LSTM framework adopts the most intuitive and direct concatenation fusion: firstly using CNN and LSTM to extract and fuse image and text features respectively, then using an FCN (fully convolutional network) to obtain the final segmentation.

In 2016, Hu et al. [17] proposed the SNLE (Segmentation from Natural Language Expression). At the initial stage, image features are extracted using FCN, and a normalized (x, y) relative coordinates channels are added to each feature pixel to represent its spatial position. After processing the text using an embedding mapping, the text is input into LSTM, and the output of the hidden layer after inputting the whole text is exerted as the text feature. Subsequently, SNLE broadcasts the text features to each pixel of the features, concatenates the coordinates directly, and afterwards feeds the concatenated features into an FCN. The response map returned by the network is upsampled and filtered to obtain the final segmentation.

However, the LSTM-extracted linguistic features cannot effectively get associated with visual features. Liu et al. [18] pointed out that when humans perform RIS, they shift their minds frequently between the image and the text, as well as consider context going back and forth. In order to simulate this approach in computer learning, Liu et al. proposed the RMI (Recurrent Multimodal Interaction) model, which concatenates the features and images extracted by LSTM each time a word is input. The concatenation results are input into the convLSTM network, and based on the output of convLSTM, the model finally performs upsampling to obtain segmentation.

Margfoy-Tuay et al. further proposed [19] the DMN (Dynamic Multimodal Network). DMN uses SRU to extract the linguistic features and convolves it with visual features to obtain a response map. At the final stage, the textual and linguistic features, response maps, and normalized spatial coordinates are fed into the multimodal SRU to generate the segmentation.

3.2. Attention-Mechanism-based Methods

Although the direct fusion method is simple and effective, when encountering complex syntax structures, the language model under the RNN framework may find it difficult to determine the dependency and importance of contextual information, which may result in unwanted comprehension bias. In 2017, Ashish Vaswani et al. of Google Brain proposed the Attention Mechanism [20], which allows neural networks to automatically learn and selectively focus on important information in inputs, and establish associations between information features from different modalities, thus enabling computers to process semantically rich texts and images.

In 2018, Yu et al. [21] proposed the MAttNet (Modular Attention Network). MAttNet segments images using MaskRCNN, and then extracts text features by both Bi-LSTM and attention mechanism to obtain features in aspects of appearance, position, and relationship between instances. These features are then combined with image features to generate segmentation.

In 2019, Ye et al. [22] proposed the CMSA (Cross Modal Self-Attention Network). CMSA first establishes a cross modal feature by fusing each image pixel p and each description word n as a (p, n) pair jointed with relative spatial coordinates, then calculates the direct correlation value of each (p, n) pair through Query and Key, and finally takes the average over the words for p to arrive at the attention representation of the cross-modal feature.

In 2020, Hu et al. [23] proposed the BCAM (Bi-directional Cross modal Attention Module). BCAM consists of two attention mechanisms: VLA (Vision Guided Linguistic Attention) and LVA (Language Guided Visual Attention). Specifically, this model first fuses image, text features, and spatial coordinates as feature units and inputs them into VLA to obtain text features with attention.

Then, the output of VLA and the feature units are input into LVA to obtain image features with attention. Finally, segmentation is obtained through convolution and upsampling.

In the same year, Hui et al. [24] proposed the LSCM (Linguistic Structure Guided Context Modeling). In LSCM, the RE input is abstracted into a graph structure, with each word being one of its node. LSCM first integrates image and text information, and then generates multimodal features, combining the image again, for each node using a cross modal attention mechanism. Finally, the Dependency Parsing Tree technique is used to analyze node correlation, adjust edge weights in the graph, and at last associated with multimodal features to obtain segmentation.

Although attention mechanisms can effectively handle the direct relationship between multimodal information, commonly used RIS datasets do not provide attention annotation information, which may lead to latent errors.

3.3. Recurrent Refinement Methods

Typically, when downsampling strategies such as pooling is applied in CNNs, the graph size is compressed exponentially, the extracted features of different scales are collectively referred to as pyramidal features, or hierarchical features. During the downsampling process, the contours and details of the graphics will be lost, making it difficult to generate accurate segmentation using only high-dimensional features of deep networks. To tackle with information loss, the model can utilize all the information from pyramidal features to recurrently refine the segmentation, as shown in Fig 2.

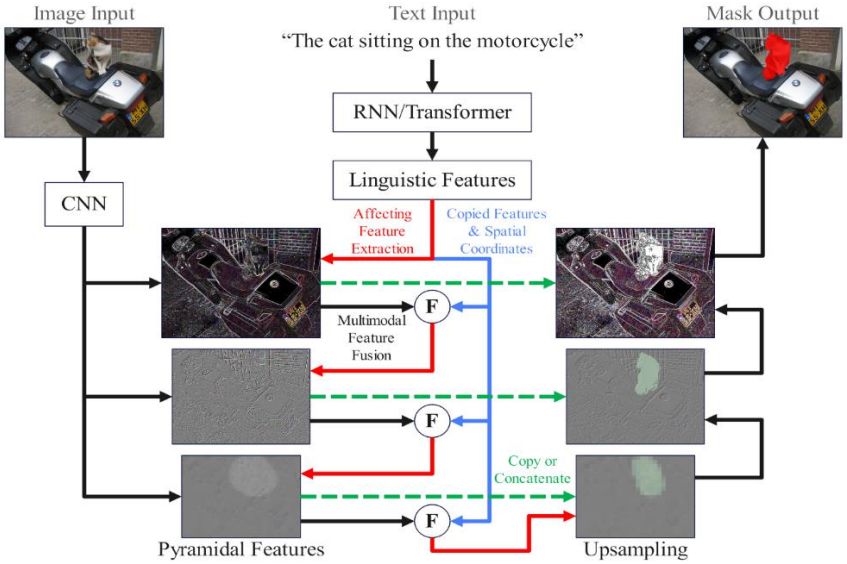

Fig 2. An illustration of a possible recurrent refinement method.

In 2018, Li et al. [25] proposed the RRN (Recurrent Refinement Network). RRN first uses the CNN-LSTM framework to fuse multimodal features, and then takes the fusion result as the initial input of convLSTM. The hierarchical image features are sequentially input into convLSTM successively from small to large scales to obtain segmentation.

In 2020, Ye et al. [26] proposed the DCLN (Dual Convolutional LSTM Network). DCLN first extracts linguistic features through Bi-LSTM and fuses to generate pyramidal features. Then, the text embedding processed by attention mechanism is multiplied with the aforementioned pyramidal features, and the segmentation is obtained through convLSTM recurrently.

In 2022, Zhao et al. [27] proposed the LAVT (Language Aware Vision Transformer). LAVT uses BERT model to extract linguistic features and fuses them with visual features pixel-wise through an attention mechanism. Features processed with attention module enter the next downsampling through a Language Gate similar to the gate structure in LSTM, hence text information can be merged into pyramid features adequately. The final segmentation is achieved by recurrently mapping the combination of small-scale visual features tand features with attention to large-scale features.

3.4. CLIP-driven Method

Based on attention mechanism, in 2021, Alec Radford et al. [28] proposed the CLIP (Comparative Language Image Pre training). CLIP mainly learns the relevance between text and images by encoding them to a unified vector space and calculating their cosine similarity. CLIP has attained feature representations of a wide range of images and texts through large-scale visual-language pre-training, providing strong support for multimodal tasks such as RIS.

In 2022, Wang et al. [29] proposed the CRIS (CLIP Driven Referring Image Segmentation). CRIS first extracts text features by Transformer and pyramidal visual features by ResNet, and then incorporates text features into image features through attention mechanism. Being passed to a Visual Language Decoder, the semantic information is propagated to each pixels of the image features. Similar to CLIP, CRIS uses Text-to-Pixel contrastive learning to maximize the similarity between text features and corresponding pixel features, and applies CLIP's zero-shot classification functionality to achieve a fine-grained segmentation.

3.5. Polygon Segmentation Method

The classical image segmentation models use a pixel-level Boolean matrix for segmentation. However, in the data used for training, the boundaries of instances are represented by a polygon enclosed by a series of coordinates, and the model needs to convert these coordinates into a Boolean matrix of that polygon as the output segmentation. Nevertheless, the original polygon representation conveys much more abundant information on instance edges, and allows models to bypass upsampling. Therefore, training the model to directly predict polygon segmentation can be an effective RIS method.

In 2023, Liu et al. [30] proposed the PolyFormer (PF). PolyFormer first uses Swin Transformer and BERT as the image and text encoder respectively, then uses a Multi-modal Transformer Encoder to fuse the features, and inputs them into the Regression-based Transformer Decoder. The Decoder loops as a usual Transformer does, outputs based on inputs a series of coordinate values as polygon segmentation.

4. Evaluation of Existing RIS Methods

Performance of RIS methods mentioned in Sec 3. on RefCOCO(UNC), RefCOCO+(UNC), RefCOCOg(UMD), RefCOCOg(Google) and ReferIt is shown in Table 2.

Table 2. Comparison of performance with state-of-the-art methods in terms of overall IoU on 4 benchmark datasets. RefCOCOg refers to both split by Google (test is invalid) and by UMD (both val and test are valid). MAN: MAttNet. DF: Direct Fusion. AM: Attention Module. RR: Recurrent Refinement. PS: Polygon Segmentation.

Method	Class	Text Encoder	RefCOCO(UNC)			RefCOCO+(UNC)			RefCOCOg		ReferIt
			val	testA	testB	val	testA	testB	val	test	test
SNLE[17]	DF	LSTM	-	-	-	-	-	-	28.14	-	48.03
RMI[18]	DF	LSTM	45.18	45.69	45.57	29.86	30.48	29.50	34.52	-	58.73
DMN[19]	DF	SRU	49.78	54.84	45.20	38.88	44.25	32.49	37.64	-	52.81
MAN[21]	AM	BiLSTM	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-
CSMA[22]	AM	-	58.32	60.61	55.09	43.76	47.60	37.89	39.98	-	63.80
BCAM[23]	AM	LSTM	61.35	63.37	59.57	48.57	52.87	42.13	48.04	-	63.46
LSCM[24]	AM	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	48.05	-	66.57
RRN[25]	RR	LSTM	55.33	57.26	53.95	39.75	42.15	36.11	36.45	-	63.63
DCLN[26]	RR	BiLSTM	59.04	60.74	56.73	44.54	47.92	39.73	41.77	-	63.92
LAVT[27]	RR	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	-
CRIS[29]	CLIP	GPT-2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36	-
PF[30]	PS	BERT	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17	67.22

5. Conclusion

This paper presents a review of a few works on Referring Image Segmentation. The paper includes introduction to the common datasets and evaluation metrics in RIS. Based on how multimodal information is encoded, fused and decoded, this paper categorized the existing RIS methods into 5 categories The research of RIS is still thriving and expectable with a promising future that is not limited by the following aspects.

Higher interpretability. Although great effort and improvement has been made in RIS, the model is still a black box with internal incomprehensible process. To better learn how we can improve the model mathematically and programmatically, higher interpretability is required.

Higher efficiency and lightweighted network structures. With the popularity of smart devices such as smartphones, cameras, etc., the design of network structures needs to be lightweighted as to meet the need of running on low-power, low-latency edge devices.

Generalized applications. RIS, as a novel technique, is yet to be applied in more domains, such as unsupervised learning like Zero-shot RIS in the case that the presented classes of objects not included in human knowledge, tracking instance in one-to-many inputs in a video clip, 3D model RIS and more unlimited applications.

References

[1] Linder, Jason, Gierad Laput, Mira Dontcheva, Gregg Wilensky, W. Chang, Aseem Agarwala and Eytan Adar. 'PixelTone: a multimodal interface for image editing.' Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2013, 2829-2830.

[2] Cheng, Ming-Ming, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J. Mitra, and Philip Torr. 'ImageSpirit: Verbal Guided Image Parsing'. ACM Transactions on Graphics, 2014, 34(1): 1–11.

[3] Wu, Dongming, Xingping Dong, Ling Shao, and Jianbing Shen. 'Multi-Level Representation Learning With Semantic Alignment for Referring Video Object Segmentation'. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 4996–5005.

[4] Kazemzadeh, Sahar, Vicente Ordonez, Mark Matten, and Tamara Berg. 'ReferItGame: Referring to Objects in Photographs of Natural Scenes'. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, 787-798.

[5] Grubinger, Michael, Paul D. Clough, Henning Müller, and Thomas Deselaers. 'The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems', 2006.

[6] Escalante, Hugo Jair, Carlos A. Hernández, Jesus A. Gonzalez, Aurelio López-López, Manuel Montes-y-Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor-Pineda, and Michael Grubinger. 'The Segmented and Annotated IAPR TC-12 Benchmark'. Comput. Vis. Image Underst. 2010, 114: 419–428.

[7] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 'Microsoft COCO: Common Objects in Context'. In Computer Vision ECCV 2014, 740–55.

[8] Yu, Licheng, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 'Modeling Context in Referring Expressions'. ArXiv abs/1608.00272 (2016).

[9] Mao, Junhua, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 'Generation and Comprehension of Unambiguous Object Descriptions'. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] Vries, Harm de, Florian Strub, A. P. Sarath Chandar, Olivier Pietquin, H. Larochelle, and Aaron C. Courville. 'GuessWhat?! Visual Object Discovery through Multi-Modal Dialogue'. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 4466–75.

[11] Liu, Chang, Henghui Ding, and Xudong Jiang. 'GRES: Generalized Referring Expression Segmentation'. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 23592–601, 2023.

[12] Yu, Jiahui, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. 'UnitBox: An Advanced Object Detection Network'. In Proceedings of the 24th ACM International Conference on Multimedia. MM '16. ACM, 2016.

[13] Rezatofighi, Seyed Hamid, Nathan Tsoi, Junyoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 'Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regre-ssion'. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 658-66.

[14] Zheng, Zhaohui, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, Dongwei Ren. ‘Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression’. In AAAI Conference on Artificial Intelligence, 2019.

[15] Salehi, Seyed Sadegh Mohseni, Deniz Erdoğmuş, and Ali Gholipour. ‘Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks’. In MLMI@MICCAI, 2017.

[16] Jun, He, Caiqing Zhang, Xiaozhen Li, Dehai Zhang. Survey of Research on Multimodal Fusion Technology for Deep Learning. Computer Engineering, 2020, 46(5): 1-11.

[17] Hu, Ronghang, Marcus Rohrbach, and Trevor Darrell. ‘Segmentation from Natural Language Expressions’. ArXiv abs/1603.06180 (2016).

[18] Liu, Chenxi, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. ‘Recurrent Multimodal Interaction for Referring Image Segmentation’. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[19] Margffoy-Tuay, Edgar, Juan C. Perez, Emilio Botero, and Pablo Arbelaez. ‘Dynamic Multimodal Instance Segmentation Guided by Natural Language Queries’. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[20] Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. ‘Attention Is All You Need’. In Neural Information Processing Systems, 2017.

[21] Yu, Licheng, Zhe L. Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. ‘MAttNet: Modular Attention Network for Referring Expression Comprehension’. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 1307–15.

[22] Ye, Linwei, Mrigank Rochan, Zhi Liu, and Yang Wang. ‘Cross-Modal Self-Attention Network for Referring Image Segmentation’. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[23] Hu, Zhiwei, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. ‘Bi-Directional Relationship Inferring Network for Referring Image Segmentation’. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 4423–4432.

[24] Hui, Tianrui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. ‘Linguistic Structure Guided Context Modeling for Referring Image Segmentation’. In European Conference on Computer Vision, 2020.

[25] Li, Ruiyu, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. ‘Referring Image Segmentation via Recurrent Refinement Networks’. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 5745–5753.

[26] Ye, Linwei, Zhi Liu, and Yang Wang. ‘Dual Convolutional LSTM Network for Referring Image Segmentation’. IEEE Transactions on Multimedia 22, 2020, 12: 3224–3235.

[27] Yang, Zhao, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. ‘LAVT: Language-Aware Vision Transformer for Referring Image Segmentation’. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 18155–18165.

[28] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. ‘Learning Transferable Visual Models From Natural Language Supervision’. In International Conference on Machine Learning, 2021.

[29] Wang, Zhaoqing, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. ‘CRIS: CLIP-Driven Referring Image Segmentation’. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 11686–11695.

[30] Liu, Jiang, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. ‘PolyFormer: Referring Image Segmentation As Sequential Polygon Generation’. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 18653–18663.