Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13413-x

A closer look at referring expressions for video object segmentation

Miriam Bellver[1] · Carles Ventura[2] · Carina Silberer[3] · Ioannis Kazakos[4] · Jordi Torres[1] · Xavier Giro-i-Nieto[5,6]

Received: 4 December 2020 / Revised: 11 March 2022 / Accepted: 2 July 2022

Abstract
The task of Language-guided Video Object Segmentation (LVOS) aims at generating binary masks for an object referred by a linguistic expression. When this expression unambiguously describes an object in the scene, it is named *referring expression* (RE). Our work argues that existing benchmarks used for LVOS are mainly composed of trivial cases, in which referents can be identified with simple phrases. Our analysis relies on a new categorization of the referring expressions in the DAVIS-2017 and Actor-Action datasets into trivial and non-trivial REs, where the non-trivial REs are further annotated with seven RE semantic categories. We leverage these data to analyze the performance of RefVOS, a novel neural network that obtains competitive results for the task of language-guided image segmentation and state of the art results for LVOS. Our study indicates that the major challenges for the task are related to understanding motion and static actions.

Keywords Referring expressions · Video object segmentation · Vision and language

1 Introduction

Video Object Segmentation (VOS) [33, 44] has been traditionally considered on setups where a user would annotate the pixels of an object in a video frame, and an automatic system would extend this to the rest of video frames where the object is visible. Our work aims at simplifying the human-computer interaction by allowing linguistic expressions as initialization cues, instead of user interactive segmentations under the form of a detailed binary mask, bounding box, scribble or points. In particular, we focus on *referring expressions (REs)*, which allow the identification of an individual object in a discourse or scene (the *referent*). REs unambiguously identify the target instance. For example, Fig. 1 depicts REs related to one of the objects contained in a video sequence, which is highlighted in green.

✉ Carles Ventura
cventuraroy@uoc.edu

Extended author information available on the last page of the article.

Fig. 1 Video sequences for DAVIS-2017 with REs and our results. The first column shows a reference frame, the second to third columns depict the masks produced by our model when given the RE shown on top. Finally, the fourth to fifth columns show the results for the REs shown on top of these columns, which refers to another object of the video sequence

Language-guided Video Object Segmentation (LVOS) was first addressed by Khoreva et al. [21], and later tackled by Gavrilyuk et al. [12] and Wang et al. [40]. Compared to related works on still images [4, 48], REs for video objects may be more complex, as they can refer to variations in the properties of the objects, such as a change of location or appearance. The particularities of REs for videos were initially addressed by Khoreva et al. [21], who built a dataset of REs divided in two categories: REs for the first frame of a video, and REs for the full clip. Our work proposes another approach for analyzing the performance of the state of the art in LVOS. We identify seven categories of REs, shown in Table 1, and use them to annotate existing datasets.

The main goal of our work is to study the task of LVOS, by focusing on the effect the different categories of REs have on the model performance in current benchmarks. We address both the language-guided image segmentation and the LVOS tasks with *RefVOS*, our end-to-end deep neural network that leverages the language representation model BERT [10] to encode the phrases into distributed representations. *RefVOS* stands for **Ref**erring Expressions for **V**ideo **O**bject **S**egmentation. Our model achieves results comparable to the state of the art for the RefCOCO dataset of still images [20], and improves the state of the art over the DAVIS-2017 [34] and Actor-Action datasets (A2D) [43] for video datasets

Table 1 The semantic categories used for annotation

Category	Q: Does RE tell you about referent r...	Example
appearance	how r looks like?	...*in a yellow dress*...
category	r's name or category (noun)	...*seagull*...
location	where r is located? (rel. to image/other object)	...*near tractor*...
motion	if r moves or changes its location?	...*walking*...
obj-motion	if r moves or changes another object's location?	...*riding a bike*...
static	what r is doing (if not moving)?	...*eating*...
obj-static	if r acts on another object (no motion)?	...*holding a bike*...

augmented with the phrases collected by Khoreva et al. [21] and Gavrilyuk et al. [12], respectively. We also identify the categories of REs which are most challenging for *RefVOS*, using our own REs annotations for the A2D dataset.

Our main contributions are summarized as follows: (1) an end-to-end model, namely *RefVOS*, that achieves state of the art performance with available expressions for DAVIS-2017 and A2D benchmarks, (2) a novel categorization of REs tailored to the video scenario with an analysis of the current benchmarks, and (3) an extension of A2D with additional REs of varying semantic information to analyze the limitations and strengths of our model according to the proposed linguistic categories.

The models, code and extended dataset of REs are available at https://github.com/miriambellver/refvos.

2 Related work

2.1 Language-guided image segmentation

The task of Language-guided Image Segmentation, also known as referring image segmentation, was first tackled by Hu et al. [16]. They used VGG-16 [39] to obtain a visual representation of the image, and a Long Short-Term Memory (LSTM) network to obtain an embedding of the RE. From the concatenation of visual and language features, the segmentation of the referred object is obtained. Posterior work [26] explored how to include multi-scale semantics in the pipeline, by proposing a Recurrent Refinement Network that takes pyramidal features and refines the segmentation masks progressively. Liu et al. [27] argued to better represent the multi-modality of the task by jointly modeling the language and the image with a multi-modal LSTM that encodes the sequential interactions between words, visual features and the spatial information. With the same purpose of better capturing the multi-modal nature of this task, long-range correlations between the visual and language representations were reinforced by learning a cross-modal attention module (CMSA) [46]. Building on the same idea, BRINet [17] added a gated bidirectional fusion module to better integrate multi-level features. STEP [4]. learned a visual-textual co-embedding that iteratively refines the textual embedding of the RE with a Convolutional Recurrent Neural Network, in a collaborative learning setup to improve the segmentation. An alternative may consist of using off-the-shelf object detectors, like MAttNet [48]. In this case, a language attention network decomposed REs into three components: subject, location, and relationships, and merged the features obtained for each into single phrase embeddings. Given the object candidate by the off-the-shelf object detector model and a RE, the visual module dynamically weighted scores from all three modules to fuse them. A different approach was proposed in CMPC [18], which leveraged multi-modal graph reasoning to identify the target objects.

Whereas previous works mainly focus on how to better exploit the REs by designing language encoders tailored to the task, our proposed architecture RefVOS leverages BERT [10] to obtain the language representations. BERT is a bidirectional language encoder based on transformers, a type of neural network architecture that is used by current state of the art models in natural language processing [35] and are also more and more used in computer vision [2, 42]. BERT is thus a strong baseline for obtaining language embeddings. Another characteristic of our work, compared to previous methods such as MAttNet [48], is that RefVOS directly produces pixel-wise annotations, without requiring object detections, which allows to train our model end-to-end. As a visual encoder, we use DeepLabv3

[5], a state-of-the-art model for segmentation. Compared to other works which perform early-fusion of multi-modal features (CMPC [18]), we apply late-fusion of the visual and language features in order to predict the final segmentation of the referred region. Thanks to late-fusion, we can work with off-the-shelf models for both the visual and the language encoders. As a conclusion, RefVOS is a simpler model trained end-to-end that obtains a performance comparable to the state of the art on still images.

2.2 Language-guided video object tracking

Object Tracking is a task similar to Video Object Segmentation as it also follows a referent across video frames, but in the tracking case the model localizes the object with a bounding box instead of a binary mask. Li et al. [24] and Feng et al. [11] tackle the object tracking problem given a linguistic expression instead of providing a bounding box for the first frame.

Our work provides pixel-wise segmentation masks that could be easily converted into bounding boxes, and at the same time avoid the annotation ambiguities that result from overlapping bounding boxes.

2.3 Language-guided video object segmentation (LVOS)

Video Object Segmentation (VOS) [33, 44] has traditionally focused on semi-supervised setups in which a binary mask of the object is provided for the first frame of the video. Khoreva et al. [21] proposed to replace the mask supervision with a linguistic expression. In their work, they extended the DAVIS-2017 dataset [34], a popular dataset for VOS, by collecting REs for the annotated objects. They provide two different kinds of annotations collected by two annotators: *first frame* annotations are the ones that are produced by only looking at the first frame of the video, whereas *full video* annotations are produced after seeing the whole video sequence. They used the image-based MAttNet [48] model pre-trained on RefCOCO to ground the localization of the referred object, and then trained a segmentation network with DAVIS-2017 to produce the pixel-wise prediction. To ensure coherent bounding boxes across frames, they enforced temporal consistency with a post-processing step. To the authors' knowledge, Khoreva et al. [21] is the only work prior to ours that focuses on REs for VOS. Related work by Gavrilyuk et al. [12] addresses a similar task by segmenting video objects given an expression. They extend the Actor-Action Dataset (A2D) [43] by collecting linguistic expressions, but some of them may be ambiguous with respect to the intended referent, as they were not produced with the aim of unique reference, but description. The authors propose a model with a 3D convolutional encoder and dynamic filters that specialize in localizing the target objects. Wang et al. [40] also leveraged 3D convolutional networks, adding cross-attention between the visual and the language encoder. Concurrent to our work, Seo et al. [38] proposed URVOS, a model for LVOS composed of a cross-modal attention module for the visual and language features, and a memory attention module to leverage information from past predictions in a sequence.

Compared to the aforementioned methods, our work proposes a simpler model trained end-to-end that treats each video frame independently. Hence, RefVOS does not exploit the temporal information, yet it outperforms all previous works. Our method has the advantage that it can be used on either still images or video sequences, and it does not rely on heavy 3D encoders such as the architectures by Gavrilyuk et al. [12] and Wang et al. [40]. In contrast to Khoreva et al.'s [21] approach, our network can be trained end-to-end as it does not rely on any detection network or post-processing step. Finally, the most recent work

URVOS [38] is a complex architecture due to the cross-attention and memory network. RefVOS proposes a much simpler architecture composed of two independent encoders with late-fusion of embeddings.

2.4 Categorization of referring expressions

RefCOCO, RefCOCO+ [49] and RefCOCOg [30] are datasets that provide REs for the still images in the MSCOCO dataset [25]. Each dataset focuses on different aspects related to the difficulty of REs: those of RefCOCO and RefCOCO+ were collected using the interactive ReferIt two-player game [20], designed to crowdsource expressions that uniquely identify the target referents. However, for RefCOCO+, *location* information was disallowed. RefCOCOg, in turn, collected non-interactively, only contains *non-trivial* instances of target objects, that is, there is at least one other object of the same class in an image (e.g., multiple *dogs*). The CLEVR [19] and CLEVRER [47] datasets contain objects of certain shapes, attributes such as sizes and colors, and spatial relationships. CLEVR uses synthetic images and linguistic expressions designed to test visual question answering systems, while our work focuses on human-produced language and natural videos. CLEVR-Ref+ [29] extends the CLEVR dataset by adding REs for Language-guided Image Segmentation instead of visual question answering, but using the same synthetic images.

Khoreva et al. [21] categorize the REs they collected for DAVIS-2017 in order to analyze the effectiveness of their proposed model. This is similar to our work, however, while they distinguish REs according to their length and whether they contain spatial words (e.g., *left*) or verbs, we propose a more fine-grained, semantic categorization, presented in Table 1, that also distinguishes between different aspects of verb meaning related to motion and object relations. Khoreva et al. [21] further analyze the REs in DAVIS-2017 with respect to the parts of speech they contain, while we use our *semantic* categories for dataset analysis.

3 RefVOS model

The task of language-guided image segmentation is to, given a still image and a linguistic expression, segment the region to which the expression refers. Language-guided video object segmentation (LVOS) is a natural extension of this task, which aims at segmenting the referred object in the different video frames of the sequence. The latter is a more complicated task as the linguistic expression can refer to the motion or scene changes in the video.

In our work, we address the task of language-guided image segmentation and LVOS with the deep neural network depicted in Fig. 2, that we call RefVOS. RefVOS operates at the frame level, i.e., it treats each frame independently, and is thus applicable to both images and videos. It uses state of the art visual and linguistic feature extractors, which are combined into a multi-modal embedding decoder to generate a binary mask for the referent.

In order to define the architecture of RefVOS, we first formalize the LVOS task: In this task, the input of the system (that we define as X) consists of an input video (V) and a linguistic expression (L), i.e., $X = \{V, L\}$. Each video sequence is composed of T video frames $V = \{I_0, I_1, I_2, \ldots I_{T-1}\}$, and each linguistic expression is composed of M tokens $L = \{w_0, w_1, w_2, \ldots w_{M-1}\}$. Note that only video frame I_t is different at each time step t, whereas the expression L is kept constant for all video frames.

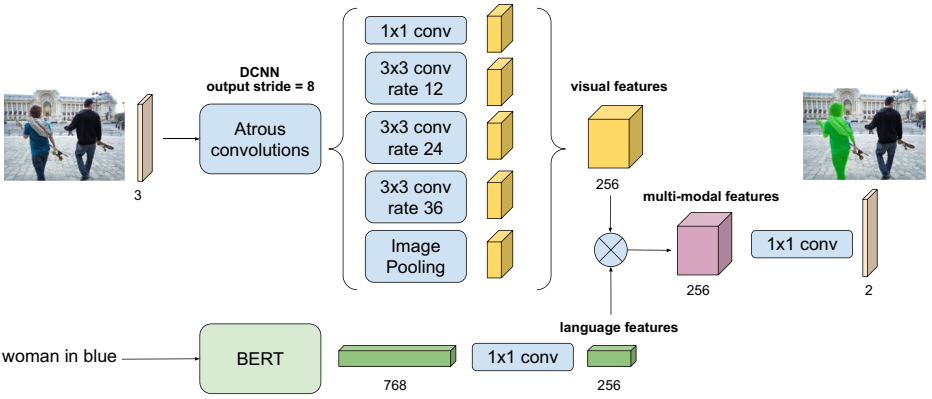

Fig. 2 Architecture of our RefVOS model

3.1 Visual encoder

To encode the images, we rely on DeepLabv3, a network for semantic segmentation based on dilate / atrous convolutions [5]. We use DeepLabv3 with a ResNet101 [14] backbone and an output stride of 8. The Atrous Spatial Pyramid Pooling (ASPP) has atrous convolutions with rates of 12, 24 and 36. Following the formulation defined in Section 3, the visual embedding is obtained independently for each video frame (I_t). In the case of the LVOS task, for each video frame $I_t \in V$, we define the visual embedding as $e_{v,t} = V_{enc}(I_t)$, where $V_{enc}()$ is the DeepLabv3 network.

3.2 Language encoder

In contrast to previous works addressing language-guided image segmentation, we are the first ones to leverage the bidirectional transformer model BERT [10] as language encoder. We use BERT in our pipeline in order to obtain an embedding for the linguistic expressions. We choose the BERT-base architecture, which is composed of 12 layers and a total of 110M parameters. We first fine-tune BERT (originally pre-trained on the BookCorpus [52] and the English Wikipedia) with the REs of RefCOCO using the masked language modelling (MLM) loss for one epoch, which consists of randomly masking a percentage of input tokens and then predicting them, following the common fine-tuning procedure for BERT. We then integrate BERT into our pipeline and fine-tune it specifically towards the language-guided image segmentation task: to this end we tokenize the linguistic expression L and add [CLS] and [SEP] tokens at its beginning and end, respectively. BERT produces a 768-dimensional embedding for each input token. We adopt the procedure of Devlin et al. [10] and extract the embedding corresponding to the [CLS] input token, i.e., the *pooled output*, as it aggregates a representation of the whole sequence. The encoded expression is then converted to a 256-dimensional embedding with a linear projection.

We formally define the output of the language encoder as $e_l = L_{enc}(L)$, where $L_{enc}()$ first tokenizes the linguistic expression L into the sequence of tokens $\{w_0, w_1, w_2, \ldots w_{M-1}\}$, then obtains the pooled output from BERT-base, and finally produces the 256-dimensional language embedding e_l. Note that this embedding is unique for the whole video sequence V.

3.3 Multi-modal embedding

At each time step t, we obtain a multi-modal embedding Y_t by performing element-wise multiplication of the language features and the visual features extracted by the ASPP from DeepLabv3. That is, $Y_t = e_{v,t} \otimes e_l$.[1] A convolutional layer then predicts two maps, one for the *foreground* and another one for the *background* class, $S_t = F(Y_t)$, where S_t is the segmentation result at time step t and F is the convolutional layer. We employ the cross-entropy loss commonly used for segmentation. As in this case there are only two class categories, the loss is the following:

$$L(S_t, \hat{S}_t) = -S_t \log \hat{S}_t - (1 - S_t) \log(1 - \hat{S}_t)$$

where $\hat{S}_t$ is the predicted segmentation and S_t is the ground truth segmentation using a binary encoding $S_t = \{0, 1\}$.

4 Referring expression categorization

We propose a novel categorization for referring expressions (REs), i.e., linguistic expressions that allow the identification of an individual object (the *referent*) in a discourse or scene. Our categorization is adapted to the challenges posed by the video object segmentation (VOS) task. We follow the commonly adopted definition of REs put forward by computational linguistics and natural language processing (e.g., [36]), and consider a (noun) phrase as a RE if it is an accurate description of the referent, but not of any other object in the current scene. Likewise, in the vision & language research field, visual RE resolution and generation has seen a rise of interest, especially in still images [8, 28, 30, 31, 50], and more recently also on videos [1, 6]. The task is formulated as, given an instance comprising an image or video with one or multiple objects, and a RE, identify the *referent* that the RE describes by predicting, e.g., its bounding box or segmentation mask. The difficulty of the task increases with the number of objects appearing in the scene, and the number of objects of the same class (e.g., multiple *dogs*). Such cases require more complex REs in order to identify the referent.

In order to make progress on VOS with REs and allow for a systematic comparison of methods, benchmark datasets need to be challenging from both the visual and linguistic perspective. However, for example, most video sequences in the DAVIS-2017 dataset used in Khoreva et al. [21] show a single object in the scene or, at most, different objects from different classes. In these cases, the actual challenge is that of predicting accurate object masks for the RE. On the other hand, the existing datasets for VOS with REs do not focus on the particularities that video information provides either, and often use object attributes which can be already captured by a single frame, or are not even true for the whole clip (e.g., the A2D dataset provides linguistic expressions for only a few frames per clip).

Our novel categorization of REs for video objects allows the analysis of datasets with respect to the *difficulty* of the REs and the kind of *semantic information* they provide. We apply it to label and analyze existing expressions of DAVIS-2017 and A2D. In addition, we use this categorization to extend a subset of the A2D test set with REs which contain semantically varying information to analyze how our model behaves with respect to the different categories.

[1] We noticed that the multiplication yielded better performance than addition or concatenation.

4.1 Difficulty and correctness of datasets

We first assess the validity and visual difficulty of a subset of DAVIS-2017 and A2D, by classifying each instance (an object and its linguistic expression) into *trivial* or *non-trivial*: if the referent is not the only object of a certain class in the video (e.g., multiple *dogs*) we consider it *non-trivial*, otherwise *trivial*. We further label each expression according to its linguistic ambiguity and correctness: we mark it as *no RE* if its referent is not the only object in the video which could be described by the expression, i.e., if it does not comply with our definition of a RE given above, and as *wrong object* if it does not match the target object of the instance but another object in the same video.

Data and Annotation Procedure Annotation was performed on the DAVIS-2017 validation set (61 REs provided by *annotator 1* [21]) in the full video setup (see Section 2), as well as on the subset of the A2D test set which contains at least two annotated objects (856 instances). Each instance contained therein was annotated by one out of four persons (all co-authors). Note that we assume the instances in A2D videos with only a single annotation as *trivial*, and automatically labeled them as such (439 instances).

Results Fig. 3 shows the proportion of expressions in the DAVIS-2017 and A2D sets with respect to their difficulty, ambiguity, and correctness. Despite being collected in a (non-interactive) referential two-player game setup, DAVIS-2017 contains a considerable proportion of ambiguous expressions (*no RE*, 8%). The proportion in A2D is slighlty higher (11%), but note that A2D was designed to contain descriptive expressions in contrast to unique references (as defined above). About 52% in DAVIS and 35% in A2D are *non-trivial* expressions, that is, more challenging for LVOS from both the linguistic and visual perspective, since the object class itself is not sufficient to uniquely identify the target object.

4.2 Semantic categorization of REs

Our categorization, shown in Table 1, is inspired by semantic categories of situations and utterances in linguistics [13, 23], tailored to the situations found in video data. Specifically, we analyze the REs with respect to the type of information they express, by assigning them categories assumed to be relevant for reference to objects in visual scenes. We focus on information relevant for both, objects in still images and videos, namely the *category*,

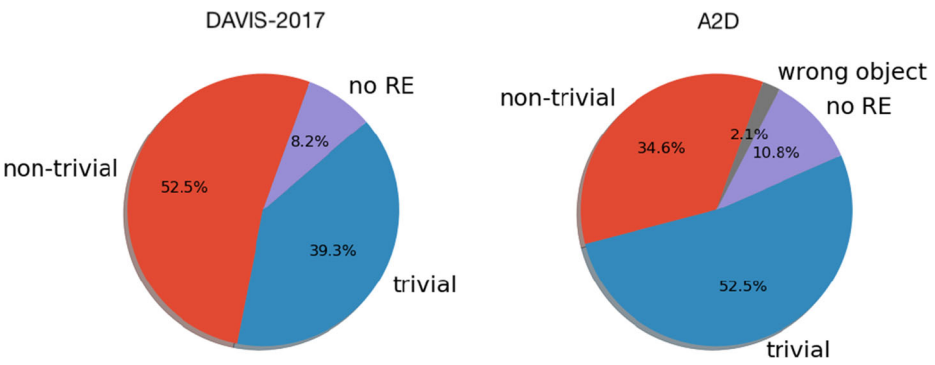

Fig. 3 Proportion of expressions in the val set of DAVIS-2017 and the test set of A2D by their difficulty ((non-)trivial), ambiguity ((no) RE), and correctness (wrong object, in A2D only)

appearance, and the *location* of the referent, and distinguish between information assumed to be more relevant for videos only, namely *motion* vs. *static* events. If, according to the RE, the referent acts upon other objects in the scene, we distinguish between whether an object is moved by the referent or not (*obj-motion* vs. *obj-static*). This information may be particularly valuable for models that reason over object interactions.

(Psycho)linguistic studies have observed a tendency of REs to contain redundant non-discriminating information, i.e., logically more information than required to establish unique reference, arguably because this reduces the effort needed for identification [15, 23]. In particular the kind (category) of the object and salient properties, such as color, have been found to be used redundantly [37]. To assess whether the phenomenon of redundancy is born out in the video datasets, we additionally label instances as *redundant* or *minimal*. A RE is labeled as *minimal* if it does not include more information than required to identify the target object, and *redundant* otherwise.

4.2.1 Data and annotation procedure

We collect annotations for the same 61 instances of the validation set of DAVIS-2017 as above, and for a subset of the test set of A2D, which we call *A2Dre* henceforth. We obtain A2Dre by selecting only instances that were labeled as *non-trivial*, which are 433 REs from 190 videos. We do not use the *trivial* cases since the analysis of such examples is not relevant, as *referents* can be described by using the *category* alone. Each annotator was presented with a RE, a video in which the target object was marked by a bounding box, and a set of questions paraphrasing our categories (see Table 1). Three annotators (all co-authors of the paper) individually labeled all instances of the DAVIS-2017 val set. The inter-annotator agreement can be considered substantial for all categories, with Davies & Fleiss' kappa coefficients [9] between $\kappa = .83$ and .97 (except *obj-static*, $\kappa = .35$, which has only 5 positively labeled instances by at most 2 annotators, and *category* which obtained perfect agreement). A2Dre was subsequently annotated by the same 3 annotators. Our final set of category annotations used for analysis was derived by means of majority voting: for each *non-trivial* RE, we kept all category labels which were assigned to the RE by at least two annotators.

4.2.2 Results: What kind of information do REs express?

First of all, we found 99% of the REs for non-trivial instances in A2Dre, and 66% in DAVIS-2017 val (74% including trivial), respectively, to contain redundant information. Recall that only the REs in DAVIS-2017 were obtained in a referential setup, the relatively larger proportion of redundant REs in A2D is therefore not surprising.

Figure 4 shows the proportion of instances in the two datasets (DAVIS-2017 val and A2Dre) that were labeled with the individual categories. As expected, the name or *category* of the referent is virtually always expressed. The visual properties of the referent, i.e., *appearance*, is prominent in both datasets, too (approx. 60%). Taken together with their high redundancy ratio, this confirms what has been found in psycholinguistic studies on reference [23]. The remaining categories, however, are rare in both datasets, or are only highly frequent in A2Dre, with *location* and *motion* being used in the majority of REs. That A2Dre comprises more complex REs than DAVIS-2017 may be not only due to their collection as descriptive, instead of discrimininative phrases, but also due to the much higher complexity of the video scenes. Note that information about referent-object interactions (*obj-static* and *obj-motion*) is neglectable, which illustrates the datasets' limited usefulness for research on

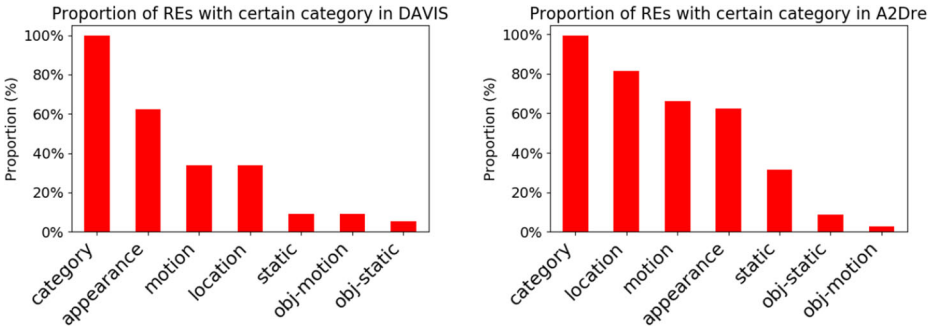

Fig. 4 REs in the validation set of DAVIS-2017 and A2Dre with respect to their categories

reasoning over object interactions [41, 45, 51]. In the experiments we report in Section 5, we discard these categories, and focus on the remaining categories only, for which we augment the A2Dre dataset.

4.3 Extending A2D with REs

As explained above, A2Dre is a subset from the A2D test set including 433 *non-trivial* REs. Due to its highly unbalanced distribution across the 7 semantic categories (Fig. 4), we select the 4 major categories *appearance, location, motion and static*. The four categories have in common that in most cases, for a given referent, a RE can be provided that expresses a certain category, and one that does not. We use these categories to augment A2Dre with additional REs, which vary according to the presence or absence of each of them. Specifically, based on our categorization of the original REs, for each RE re and category C, we produce an additional RE re' by modifying re slightly such that it does (or does not) express C. For example, for the last RE in Fig. 6, i.e. *girl in yellow dress standing near the woman*, which could be categorized as *appearance* (App+), *location* (Loc+), no *motion* (Motion-) and *static* (Static+), we produce new REs for each category: *girl standing near the woman* (no *appearance*, App-), *girl in yellow dress standing* (no *location*, Loc-), *girl in yellow dress walking* (*motion*, Motion+) and *girl in yellow dress near the woman* (no *static*, Static-). We do not apply this procedure for *category*, since it is expressed in almost all REs, and its removal may be difficult in many cases. We will refer to this extended dataset as A2Dre+.

5 Experiments

We report results with our model on two different tasks: *language-guided image segmentation* and *language-guided video object segmentation (LVOS)*. The results for still images are obtained on RefCOCO and RefCOCO+ [49], while those for videos correspond to DAVIS-2017 and A2D.

5.1 Language-guided image segmentation

Table 2 shows the impact of BERT embeddings in our model on both RefCOCO and RefCOCO+, compared with a bidirectional LSTM similar to Chen et al. [4] for encoding the linguistic expression. In particular, we average the GloVe embeddings [32] of each

Table 2 Overall IoU for RefCOCO and RefCOCO+

	RefCOCO			RefCOCO+		
	val	testA	testB	val	testA	testB
Ours with Bi-LSTM	48.46	52.90	44.43	35.35	40.72	28.43
Ours with BERT	58.65	62.28	54.28	42.07	46.46	34.23
Ours with BERT (+MLM loss)	59.45	63.19	54.17	44.71	49.73	36.17
MattNet [48]	56.51	62.37	51.70	46.67	52.39	40.08
CMSA [46]	58.32	60.61	55.09	43.76	47.60	37.89
LANG2SEG [7]	58.90	61.77	53.81	–	–	–
STEP (1-fold) [4]	56.58	58.70	55.39	–	–	–
STEP (4-fold) [4]	59.13	–	–	–	–	–
STEP (5-fold) [4]	60.04	63.46	58.97	48.18	52.33	40.41
BRINet [17]	61.35	63.37	59.57	48.57	52.87	42.13
CMPC [18]	**61.36**	**64.53**	**59.64**	**49.56**	**53.44**	**43.23**

MLM loss refers to the masked language modelling loss used training BERT with the REs from RefCOCO

token and concatenate the mean embeddings of the forward and backward pass. This baseline is compared to two configurations that use BERT. The first one fine-tunes BERT for the language-guided image segmentation task, and significantly boosts performance over using GloVe embeddings. The second has an additional step, that consists in first training BERT using the masked language modelling (MLM) loss with the REs from RefCOCO, as explained in Section 3, and then fine-tuning BERT on the language-guided image segmentation task (as in the previous configuration). We see that this configuration brings an additional gain.

Table 2 also compares our model with the state of the art on language-guided image segmentation. STEP [4] consists of an iterative model that refines the RE representation to improve the segmentation. Note that the model must be run for each iteration. Our model surpasses STEP (1-fold), which corresponds to a comparable computational cost, on RefCOCO val and testA, and is still slightly better than STEP (4-fold). Compared to STEP (5-fold), the performance of our method is slightly lower. BRINet [17] and CMPC [18] are both superior in terms of performance. However, compared to ours, they are significantly more complex. CMPC is composed of several independent modules and needs to build a relational graph per query. BRINet has a cross-attentional and a bidirectional module to fuse cross-modal features. Both BRINet and CPMC use a Dense-CRF post-processing step [22]. In comparison, our network is simpler and is fully end-to-end trainable. Qualitative results generated with our best model on RefCOCO are depicted in Fig. 5. We note how our model distinguishes properly the referred instance and generates an accurate mask. We conclude that our approach is competitive with the state of the art for language-guided image segmentation. Hence, *RefVOS* is a valid model for language-guided VOS, and for running an analysis on our RE categorization.

Fig. 5 Qualitative results obtained on RefCOCO

5.2 Language-guided video object segmentation (LVOS)

Our model is assessed for LVOS on DAVIS-2017 and A2D. In both cases, each video frame is treated separately, so we use the same architecture as in the experiments on still images in Section 5.1.

Our experiments on the DAVIS-2017 validation set are reported in Table 3. All models are pre-trained on RefCOCO. Results are provided with the J&F metric adopted in the DAVIS-2017 challenge for the two different types of REs collected by Khoreva et al. [21] explained in Section 2. J&F is the average between a region-based evaluation measure (J) and a contour-based one (F). Our experiments indicate that our baseline model trained only with RefCOCO already outperforms the best model by Khoreva et al. [21], despite the latter being fine-tuned on the same DAVIS-2017 dataset (+Ft DAVIS segms.). The difference increases when our model is fine-tuned with the segmentations provided in the training set, but freezing the language encoder. This is the configuration comparable to Khoreva et al. [21] in terms of training data, and brings gains of 2.7 and 4.9 points for the *first frame*

Table 3 J&F on DAVIS-2017 validation set

Model	+Ft DAVIS	+Ft DAVIS REs		J&F	
	segms.	1st frame	full video	1st frame	full video
midrule Khoreva et al. [21]	✓			39.3	37.1
URVOS [38]	✓	✓		44.1	–
				39.8	40.8
RefVOS	✓			42.0	42.0
	✓	✓		**44.5**	**45.1**
	✓		✓	42.7	**45.1**

and *full video* REs, respectively. Finally, we also fine-tune the BERT language encoder, obtaining a significant extra gain in performance. We want to highlight that our frame-based model does not rely on any post-processing to add temporal coherence, or optical flow, in contrast to Khoreva et al. [21], so our method may be more efficient computationally. We also compare our model to URVOS [38], a concurrent work to ours. RefVOS performs slightly better when trained with the same amount of annotated data. Qualitative results for full video REs are shown in Fig. 1. When the multiple objects belong to different categories, the model produces accurate masks from the language query, whereas it is more challenging to properly segment the referent in cases where there are multiple instances of the same class in the sequence (3rd row). The fine-tuning is done with the *full video* REs, and the REs shown in Fig. 1 are of the same kind. We note how the referred object is in general identified and properly segmented.

The results for A2D are shown in Table 4, using the metrics that allow us a comparison with previous works [12, 40]. Our model trained only with A2D already outperforms Gavrilyuk et al. [12] in *Precision* at a high threshold and at the *Overall* and *Mean Intersection Over Union (IoU)*. Moreover, our model significantly increases its performance when it is first trained on RefCOCO and later fine-tuned on A2D, both its visual and language branches. In this setup, it achieves state of the art results in all metrics by significant margins. Note that both Gavrilyuk et al. [12] and Wang et al. [40] leverage an encoder that was pre-trained on the Kinetics dataset, which includes 650,000 video clips [3]. Hence, these models see a large amount of annotated data for action recognition in videos. We also want to stress our higher *Precision* values at high thresholds (Prec@0.9) compared to previous works [12, 40], which indicate that our model is able to produce more accurate masks. Visualizations with our model are illustrated in Fig. 6.

In conclusion, RefVOS is state of the art for DAVIS-2017 and A2D on the LVOS task, although it is a frame-based model. This motivates the analysis of our model when tested with different types of REs, based on the categorization and difficulty analysis proposed in Section 4.

5.3 Referring expressions analysis for LVOS

Firstly, we analyze the performance on *trivial* and *non-trivial* linguistic expressions for both the A2D test and DAVIS-2017 validation sets. The *mean IoU* per referent obtained for *trivial* and *non-trivial* is 48.7 *vs.* 46.2 on DAVIS-2017, and 53.9 *vs.* 33.2 on A2D. We observe that the performance is worse for the *non-trivial* cases for both datasets as expected, with a major drop on A2D.

Table 4 *Precision*, *overall IoU* and *mean IoU* on A2D

	Prec		IoU	
	@0.5	@0.9	Overall	Mean
Gavrilyuk et al. [12]	50.0	0.4	55.1	42.6
Wang et al. [40]	55.7	2.0	60.1	49.0
RefVOS with A2D	49.5	6.4	59.9	43.0
RefVOS with RefCOCO	27.9	3.4	41.4	25.6
+ finetuned on A2D	**57.8**	**9.3**	**67.2**	**49.7**

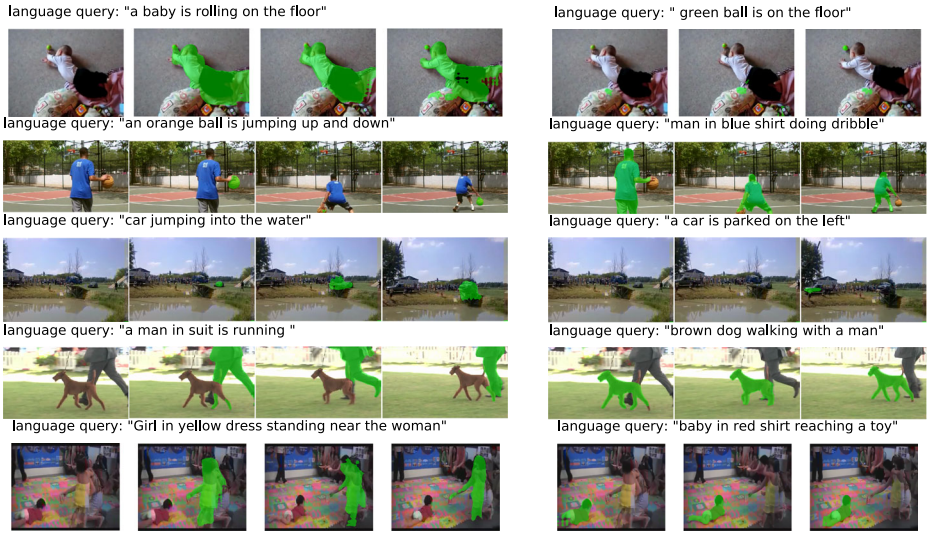

Fig. 6 Video sequences for A2D with expressions (language queries) and the results of our model. The first column shows a reference frame, the second to fourth columns depict the masks produced by our model when given the expression shown on top. Finally, the fifth to seventh columns show the results for the expression shown on top of these columns, which refers to another object of the video sequence

Secondly, we study the effect of RE categories in relation to the performance of RefVOS. The A2Dre+ dataset described in Section 4.3 allows us to have the same number of referents for all major categories: *appearance, location, motion* and *static*. Each of our referents is annotated with highly similar REs (two for each category) and are hence directly comparable. In contrast, Khoreva et al. [21] split the videos into two different subsets with non-comparable referents. Table 5 compares the performance of RefVOS depending on whether each of the categories C is present in the RE (C+, the RE expresses C) or not (C-). The results show that the presence of *appearance* (*App*+) and *location* (*Loc*+) categories yields significantly higher results compared to their absence (*App*- and *Loc*-, respectively). We also observe a drop in performance when the *static* category is present (*Static*+), which indicates that the model struggles at identifying a referent based on static actions such as *holding, sitting, eating*. In contrast, the presence (*Motion*+) or absence (*Motion*-) of the *motion* category does not affect the performance, which actually means that the model is unable to benefit from this type of REs.

In what follows, we further visually analyze the results obtained with RefVOS depending on the categories that appear in the corresponding REs. Figure 7 includes examples of the results of our model with A2Dre+. We analyze the categories *appearance*, *location*, *motion* and *static*, which are the most common semantic categories. Each column is a first frame of a video sequence with a *non-trivial* case, and each row is a different RE that has or has not a certain category. As we concluded with the quantitative results, the performance

Table 5 Effect of the presence (C+) or absence (C-) of a category C in REs

App+	App-	Loc+	Loc-	Motion+	Motion-	Static+	Static-
33.90	30.15	**34.15**	30.78	35.58	**35.60**	34.28	**36.21**

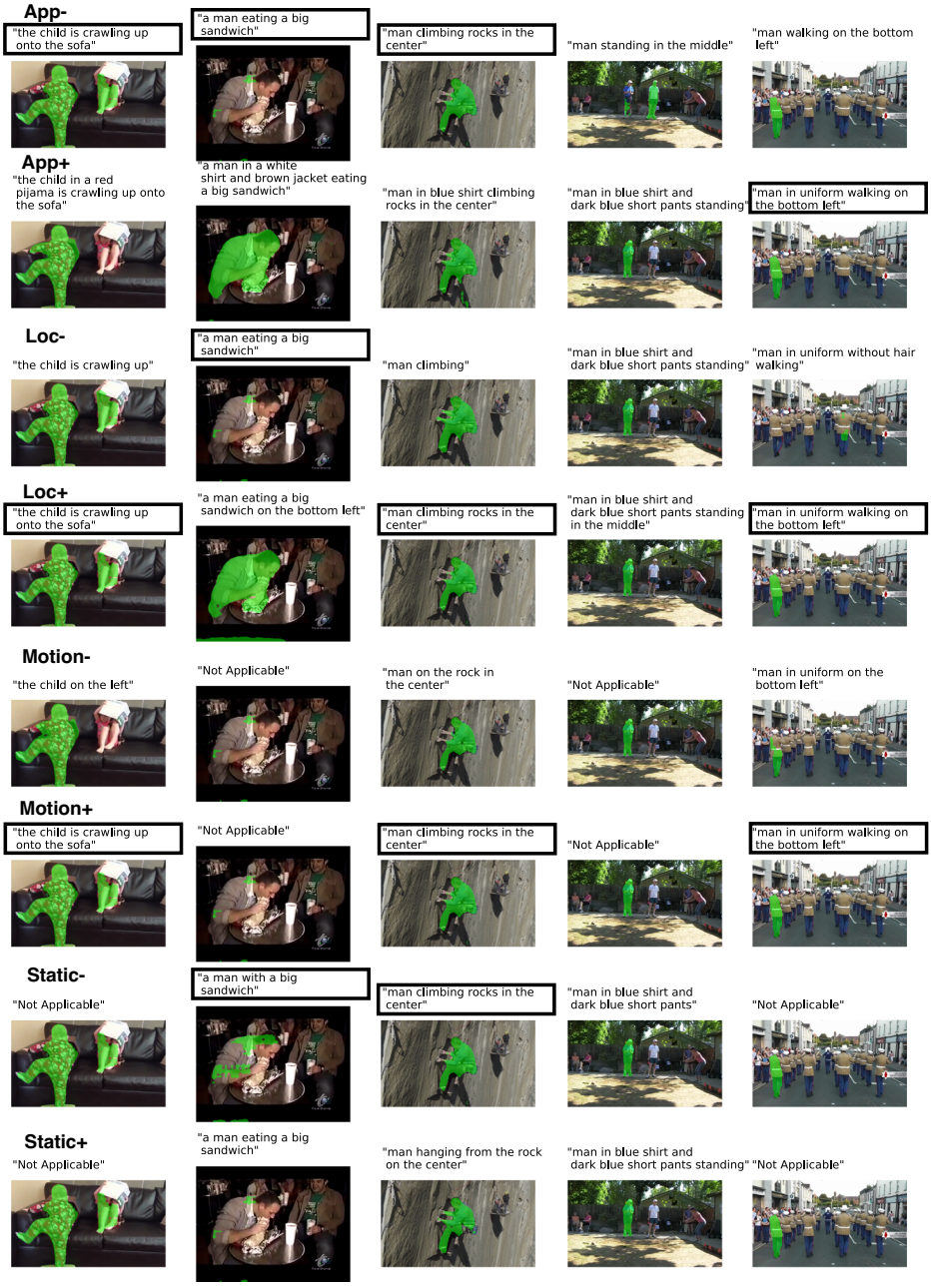

Fig. 7 Each column is the first frame of a video sequence of A2D. Each row indicates if the RE that produces the depicted result does (*C*+) or does not (*C*-) contain a certain category *C* from our proposed categorization. The framed REs are the original REs in the dataset [12]. For the example in the fourth column, the natural expression in the annotations from Gavrilyuk et al. [12] is *man standing*, which is not a RE as it does not uniquely identify the target object as defined in Section 4. For this reason, for this example all REs shown in the Figure are our own annotations

Table 6 *Overall* and *Mean IoU* on A2D for different levels of information in REs

	Overall IoU			Mean IoU		
	Trivial	Non-Trivial	All	Trivial	Non-Trivial	All
Generic	45.6	18.1	41.6	34.6	10.0	29.6
Only Actor	65.6	34.8	60.8	51.5	22.8	45.7
Only Action	56.3	30.7	52.6	43.0	18.5	38.0
Actor + Action	66.6	37.3	62.2	51.3	24.8	45.9
Full phrase	**70.2**	**47.5**	**67.2**	**53.9**	**33.2**	**49.7**

when the *appearance* and *location* categories are present is higher compared to when these categories are absent. Regarding the *motion* and *static* categories, we first notice that the annotators considered it impossible for some cases to create corresponding REs. We indicate those examples with the "Not Applicable" label. We see how the presence or the absence of the motion and static categories has a minimal impact to the results. In fact, adding these categories (through a description of the shown action, e.g., "crawling" for *Motion+* or "eating" for *Static+*) even leads to worse segmentations, like in the example "a man eating a big sandwich" from the second column in Fig. 7.

Finally, in Table 6 we study the effect of feeding the model with only the *actor*, the *action*, or the *actor and action*, without formulating any RE, for all the test set of A2D. These *actor* and *action* terms are obtained from the dataset collected by Gavrilyuk et al. [12]. In most cases these expressions are not REs as they do not unambiguously describe the referent in the video (cf. Section 4). Additionally, we consider a generic phrase *thing*. We distinguish between *trivial* and *non-trivial* cases. Results show that RefVOS works significantly better when the *actor* is provided than when the *action* is. Furthermore, performance improves when using both. Finally, having the full linguistic phrase is still the best model. Remarkably, our configuration with *actor and action* reaches higher *Overall IoU* than previous works that use complete linguistic phrases (see Table 4). Note that using the full phrase improves performance especially for the *non-trivial* cases, as these require complete linguistic expressions to identify the *referent*. We also want to stress that the aggregated performance, i.e., considering *all* cases, is dominated by the performance of the *trivial* ones, as they represent most of the dataset.

6 Conclusions

This work studies the difficulty of REs from benchmarks on Language-guided Video Object Segmentation (LVOS), and proposes seven semantic categories to analyze the nature of such REs. We introduce RefVOS, a novel model that is competitive for language-guided image segmentation, and state of the art for LVOS. However, our analysis shows that benchmarks are mainly composed of trivial cases, in which referents can be identified with simple linguistic expressions. This indicates that the reported metrics for the task may be misleading. Thus, we focus on the non-trivial cases. We extend A2D with new REs with diverse semantic categories for non-trivial cases, and test our model with them, which reveals that it struggles at exploiting motion and static events, and that it mainly benefits from REs based on appearance and location. We reckon that future research on LVOS should focus on non-trivial cases describing motion and events, as they present a challenge for language

grounding on videos. Concurrent to our work, Seo et al. [38] collected Refer-Youtube-VOS, a large-scale benchmark for LVOS built on top of Youtube-VOS [44]. We believe that, as future work, our categorization for REs could be used to classify the provided linguistic expressions by this benchmark. Thus, models could be evaluated based on the non-trivial cases and the different categories in order to analyze which REs are more challenging when using a large-scale dataset.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was partially supported by the projects PID2019-107255GB-C22 and PID2020-117142GB-I00 funded by MCIN/ AEI /10.13039/501100011033 Spanish Ministry of Science, and the grant 2017-SGR-1414 of the Government of Catalonia. This work was also partially supported by the project RTI2018-095232-B-C22 funded by the Spanish Ministry of Science, Innovation and Universities.

References

1. Anayurt H, Ozyegin SA, Cetin U, Aktas U, Kalkan S (2019) Searching for ambiguous objects in videos using relational referring expressions. In: Proceedings of the british machine vision conference (BMVC)
2. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer
3. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
4. Chen DJ, Jia S, Lo YC, Chen HT, Liu TL (2019) See-through-text grouping for referring image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 7454–7463
5. Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587
6. Chen X, Ma L, Chen J, Jie Z, Liu W, Luo J (2018) Real-time referring expression comprehension by single-stage grounding network. arXiv:1812.03426
7. Chen YW, Tsai YH, Wang T, Lin YY, Yang MH (2019) Referring expression object segmentation with caption-aware consistency. In: British machine vision conference (BMVC)
8. Cirik V, Berg-Kirkpatrick T, Morency LP (2018) Using syntax to ground referring expressions in natural images AAAI
9. Davies M, Fleiss JL (1982) Measuring agreement for multinomial data. Biometrics 38(4):1047–1051
10. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics, pp 4171–4186
11. Feng Q, Ablavsky V, Bai Q, Li G, Sclaroff S (2020) Real-time visual object tracking with natural language description. In: The IEEE winter conference on applications of computer vision, pp 700–709
12. Gavrilyuk K, Ghodrati A, Li Z, Snoek CG (2018) Actor and action video segmentation from a sentence. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5958–5966
13. Hamp B, Feldweg H (1997) GermaNet - a lexical-semantic net for German. In: Automatic information extraction and building of lexical semantic resources for NLP applications. https://www.aclweb.org/anthology/W97-0802. Accessed 20 July 2022
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
15. Hervás R, Finlayson M (2010) The prevalence of descriptive referring expressions in news and narrative. In: Proceedings of the ACL 2010 conference short papers, pp 49–54. Association for Computational Linguistics, Uppsala, Sweden. https://www.aclweb.org/anthology/P10-2010. Accessed 20 July 2022

16. Hu R, Rohrbach M, Darrell T (2016) Segmentation from natural language expressions. In: European conference on computer vision, pp 108–124. Springer
17. Hu Z, Feng G, Sun J, Zhang L, Lu H (2020) Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4424–4433
18. Huang S, Hui T, Liu S, Li G, Wei Y, Han J, Liu L, Li B (2020) Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10488–10497
19. Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R (2017) Clevr: a diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2901–2910
20. Kazemzadeh S, Ordonez V, Matten M, Berg T (2014) ReferItGame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 787–798. https://www.aclweb.org/anthology/D14-1086. Accessed 20 July 2022
21. Khoreva A, Rohrbach A, Schiele B (2018) Video object segmentation with language referring expressions. In: Asian conference on computer vision, pp 123–141. Springer
22. Krähenbühl P, Koltun V (2011) Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems, pp 109–117
23. Levelt WJM (1989) Speaking: From Intention To Articulation. MIT Press, Cambridge, MA
24. Li Z, Tao R, Gavves E, Snoek CG, Smeulders AW (2017) Tracking by natural language specification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6495–6503
25. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision – ECCV 2014, pp 740–755. Springer International Publishing. Cham
26. Liu C, Lin Z, Shen X, Yang J, Lu X, Yuille A (2017) Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1271–1280
27. Liu D, Zhang H, Wu F, Zha ZJ (2019) Learning to assemble neural module tree networks for visual grounding. In: Proceedings of the IEEE international conference on computer vision, pp 4673–4682
28. Liu J, Wang L, Yang M (2017) Referring expression generation and comprehension via attributes. In: IEEE International conference on computer vision, ICCV 2017, Venice, Italy, october 22-29, 2017, pp 4866–4874. IEEE Computer Society
29. Liu R, Liu C, Bai Y, Yuille AL (2019) Clevr-ref+: Diagnosing visual reasoning with referring expressions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4185–4194
30. Mao J, Huang J, Toshev A, Camburu O, Yuille AL, Murphy K (2016) Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 11–20
31. Nagaraja VK, Morariu VI, Davis LS (2016) Modeling context between objects for referring expression understanding. In: European conference on computer vision (ECCV)
32. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
33. Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M, Sorkine-Hornung A (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 724–732
34. Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A, Van Gool L (2017) The 2017 davis challenge on video object segmentation. arXiv:1704.00675
35. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X (2020) Pre-trained models for natural language processing: A survey. Sci China Technol Sci, pp 1–26
36. Reiter E, Dale R (1992) A fast algorithm for the generation of referring expressions. In: COLING 1992 volume 1: The 15th international conference on computational linguistics. https://www.aclweb.org/anthology/C92-1038. Accessed 20 July 2022
37. Rubio-Fernández P (2016) How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. Front Psychol 7:153
38. Seo S, Lee JY, Han B (2020) Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: Proceedings of the european conference on computer vision (ECCV)
39. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR)

40. Wang H, Deng C, Yan J, Tao D (2019) Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In: Proceedings of the IEEE international conference on computer vision, pp 3939–3948
41. Wang P, Wu Q, Cao J, Shen C, Gao L, Van Den Hengel A (2018) Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. 2019 IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 1960–1968
42. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) Segformer: Simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst, vol 34
43. Xu C, Hsieh SH, Xiong C, Corso JJ (2015) Can humans fly? action understanding with multiple classes of actors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2264–2273
44. Xu N, Yang L, Fan Y, Yang J, Yue D, Liang Y, Price B, Cohen S, Huang T (2018) Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the european conference on computer vision (ECCV), pp 585–601
45. Yang S, Li G, Yu Y (2019) Dynamic graph attention for referring expression comprehension. 2019 IEEE/CVF Int Conf Comput Vis (ICCV), pp 4643–4652
46. Ye L, Rochan M, Liu Z, Wang Y (2019) Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10502–10511
47. Yi K, Gan C, Li Y, Kohli P, Wu J, Torralba A, Tenenbaum JB (2019) Clevrer: Collision events for video representation and reasoning. In: International conference on learning representations
48. Yu L, Lin Z, Shen X, Yang J, Lu X, Bansal M, Berg TL (2018) Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1307–1315
49. Yu L, Poirson P, Yang S, Berg AC, Berg TL (2016) Modeling context in referring expressions. In: European conference on computer vision, pp 69–85. Springer
50. Yu L, Tan H, Bansal M, Berg TL (2017) A joint speaker-listener-reinforcer model for referring expressions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7282–7290
51. Zhang C, Li W, Ouyang W, Wang Q, Kim WS, Hong S (2019) Referring expression comprehension with semantic visual relationship and word mapping. In: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, pp 1258–1266. Association for Computing Machinery, Nice, France
52. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the 2015 IEEE international conference on computer vision (ICCV), ICCV '15, pp 19–27. IEEE Computer Society, USA. https://doi.org/10.1109/ICCV.2015.11

Affiliations

Miriam Bellver[1] · Carles Ventura[2] · Carina Silberer[3] · Ioannis Kazakos[4] · Jordi Torres[1] · Xavier Giro-i-Nieto[5,6]

Miriam Bellver
miriam.bellver@bsc.es

Carina Silberer
carina.silberer@ims.uni-stuttgart.de

Ioannis Kazakos
edem010@mail.ntua.gr

Jordi Torres
jordi.torres@bsc.es

Xavier Giro-i-Nieto
xavier.giro@upc.edu

1 Barcelona Supercomputing Center (BSC), Barcelona, Spain

2 Universitat Oberta de Catalunya (UOC), Barcelona, Spain

3 Institute for NLP, University of Stuttgart, Stuttgart, Germany

4 National Technical University of Athens, Athens, Greece

5 Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain

6 Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Catalonia, Spain