SynthRef: Generation of Synthetic Referring Expressions for Object Segmentation

Ioannis Kazakos[1,2]**, Carles Ventura**[3]**, Míriam Bellver**[1,4]**, Carina Silberer**[5]**, Xavier Giró-i-Nieto**[1,4]

[1]*Universitat Politècnica de Catalunya*   [2]*National Technical University of Athens*   [3]*Universitat Oberta de Catalunya*

[4]*Barcelona Supercomputing Center*   [5]*University of Stuttgart*

Abstract

Recent advances in deep learning have brought significant progress in visual grounding tasks such as language-guided video object segmentation. However, collecting large datasets for these tasks is expensive in terms of annotation time, which represents a bottleneck. To this end, we propose a novel method, namely SynthRef, for generating synthetic referring expressions for target objects in an image (or video frame), and we also present and disseminate the first large-scale dataset with synthetic referring expressions for video object segmentation. Our experiments demonstrate that by training with our synthetic referring expressions one can improve the ability of a model to generalize across different datasets, without any additional annotation cost. Moreover, our formulation allows its application to any object detection or segmentation dataset. Project site: https://github.com/imatge-upc/synthref

1   Introduction

Visual grounding tasks provide challenging benchmarks for artificial intelligence systems, as they must combine vision and language effectively. Among them, we focus on referring video object segmentation, in which a language query defines which instance to segment from a video sequence. In particular, we define *referring expressions (REs)* as linguistic phrases that allow the unique identification of an individual object (the *referent*) in a discourse or scene. (*cf.*, Reiter and Dale 1992; Qiao et al. 2020). One of the biggest challenges for this task is the lack of relatively large annotated datasets since a tremendous amount of time and human effort is required for annotation.

Using referring expressions to identify objects in the real world lies at the core of human communication. Their use for segmenting objects in images has been previously addressed (Hu et al., 2016; Liu et al., 2017; Yu et al., 2018; Ye et al., 2019; Chen et al., 2019) and has benefited from large scale datasets, such as RefCOCO (Kazemzadeh et al., 2014). However, fewer works have explored the segmentation of objects using REs in the video domain, although this provides the more natural setup compared to the image domain. Humans use referring expressions to identify objects for others in a moving world, better represented by videos than by still images. Khoreva et al. (2018) were the first to transfer the referring expression segmentation task from images to videos by collecting referring expressions for the DAVIS-2017 (Pont-Tuset et al., 2017) dataset. Later Gavrilyuk et al. (2018) provided natural language descriptions as guidance for actor segmentation in A2D (Xu et al., 2015) and J-HMDB (Jhuang et al., 2013), two datasets used for action and human pose recognition and segmentation. Finally, the first large-scale benchmark for referring video object segmentation, Refer-YouTube-VOS, was built by Seo et al. (2020) on top of YouTube-VOS (Xu et al., 2018), a benchmark for video object segmentation.

As an alternative to collecting REs from annotators, we propose generating synthetic referring expressions for an image, using only the ground truth annotations of objects and their predicted visual attributes from an off-the-shelf deep learning model. We apply this method to build a large-scale dataset with synthetic referring expressions for video object segmentation, based on an existing benchmark dataset for video instance segmentation. We use our synthetic dataset for pretraining a deep neural network for the task of referring video object segmentation and evaluate our method on two benchmark datasets used in language-guided video object segmentation.
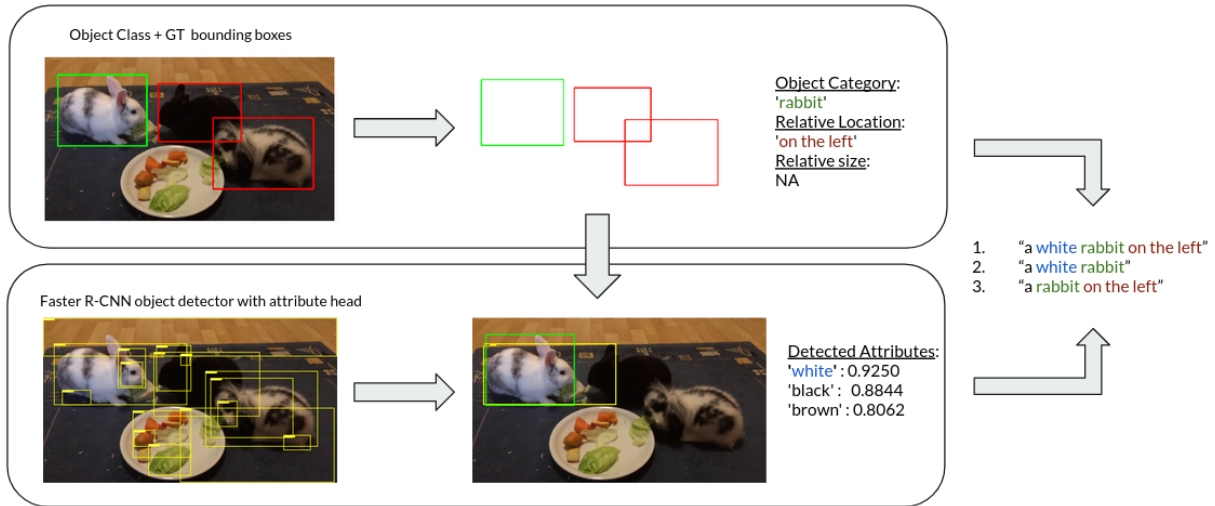
Figure 1: Overview of our method for generating synthetic referring expressions. Top: Ground truth labels (object class + bounding boxes) are used to compute a target object's relative location and size. Bottom: A Faster R-CNN object detector with attribute head predicts visual attributes for the detected objects, which are filtered by ground truth annotations. The combined cues create a set of referring expressions that uniquely describe the target object.

2 Method and Dataset

Our synthetic referring expressions are based on the ground-truth annotations of YouTube-VIS (Yang et al., 2019) dataset, described in the supplementary material. Specifically, we use the classes and bounding boxes of the target and other objects in a video frame, to determine a set of cues from which we heuristically generate a referring expression that is close to a natural language expression. We call our approach *SynthRef*, illustrated in Figure 1. We use the following four cues for generating referring expressions for a target object:

1. Object class In trivial cases where a single object of a known class is present, using the object class is enough to generate a referring expression. However, most cases involve multiple objects of the same class, thus other cues are necessary in order to disambiguate between instances.

2. Relative size If the total area of the bounding box of the referent is twice as big/small than the area(s) of the respective bounding boxes of the other object(s) of the same class, a characterization of *"bigger/smaller"* or *"the biggest/smallest"* is added to the synthetic referring expression, *e.g.* , *"the smallest dog"*.

3. Relative location In scenarios where two or three objects of the same class are present in a video frame, relative location between these objects may suffice to disambiguate between them. If the bounding boxes of the objects are fully separable, or partially above a certain threshold, then we

assume that relative location of the referent with respect to the other object(s) of the same class can be used in order to generate a non-ambiguous referring phrase. In this case, the steps for determining relative location are the following:

1. The axis which is the most separative for the bounding boxes of the two objects is determined.
2. According to the axis found and the position of the bounding boxes, a relative location description is given out of 4 options: {*"on the right", "on the left", "in the back", "in the front"*}.
3. If there are two other objects of the same class, steps 1 & 2 are computed between the referent and each of the two other objects, and the results are combined, *e.g.* , *"in the middle", "in the back right"*, *etc.* .

4. Attributes We pretrain Faster R-CNN (Ren et al., 2015) on Visual Genome (Krishna et al., 2017) for object and attribute detection (Tang et al., 2020). This model analyzed the video frames of YouTube-VIS (Yang et al., 2019) dataset to obtain, for each frame of a video, a set of detected objects (with their bounding box coordinates) and their predicted attributes. The detected bounding box with the highest overlap, in terms of Intersection-over-Union (IoU), with the ground truth bounding box of the target object is considered as the prediction corresponding to the

Dataset	Videos / Objects / Categories	Expressions	Expressions	Expressions/Object	Words/Expression
A2D Sentences	3,782 / 4,825 / 8	*Human*	6,656	1.4	7.3
DAVIS-2017	150 / 386 / 78	*Human*	1,544	4.0	5.5
Refer-YouTube-VOS	3,975 / 7,451 / 94	*Human*	27,899	3.7	7.5
SynthRef-YouTube-VIS	2,238 / 3,774 / 40	*Synthetic*	15,798	4.2	4.4

Table 1: Statistics of our dataset and comparison to existing ones. The last two columns represent the average number of unique referring expressions per object and the average number of words per referring expression respectively.

referent, as long as its IoU is over 50%. Figure 1 shows the full pipeline: Tang et al.'s (2020) model can detect a total of 201 attributes, which we group to color-like and not color-like attributes, where the latter can be both adjectives (*e.g.* , *"large"*, *"spotted"*) or verbs (*e.g.* , *"walking"*, *"surfing"*). The ones with the highest prediction score, if above a certain threshold, are selected for the two subsets, while combinations of two colors are also possible if their scores are very close (*e.g.* , *"a yellow and green parrot"*). We add an attribute to the referring expression only if no other objects belonging to the same class share the same attribute, so that the expression is able to disambiguate between instances.

Finally, we combine the aforementioned components in a natural order and add a proper article to the sentence, ending up with a synthetic referring expression. There might be cases where the generated synthetic language expression may be ambiguous, especially in cases of many similar objects of the same class, although in most cases the generated expression uniquely identifies the referent. SynthRef treats each video frame separately, so we do not force any temporal coherence of the REs. Actually, since an object may change its location or appearance throughout the video, we generate one or more synthetic referring expressions for each frame of the video. In this way, a model can be trained with different referring expressions for the same video or frame increasing its ability to generalize.

Basic statistics of our *SynthRef-YouTube-VIS* dataset and a comparison with other relevant ones are presented in Table 1. The comparison shows that our dataset, despite not being the largest one in terms of number of annotated objects and their categories, it still has the highest average number of unique referring expressions per annotated object (4.2) without involving any human annotation cost. The average number of words in our referring expressions is 4.4, which is smaller than those of

the other datasets as our goal is to generate simple and efficient synthetic referring expressions. We point out three limitations of our dataset/method: (a) the predicted attributes may be wrong or not disambiguating, (b) the relative location is not applied for more than three objects of the same class, and (c) when none of our rules can be applied, SynthRef uses just the object class (*e.g.* *"a dog"*), even if there are more instances of that class.

3 Experiments

We show the benefits of our synthetic dataset SynthRef-YouTube-VIS by using it in extending the training dataset of RefVOS (Bellver et al., 2020), a state of the art model for referring video object segmentation. The first experiments focus on DAVIS-2017 (Pont-Tuset et al., 2017), and the latter on Refer-YouTube-VOS (Seo et al., 2020).

RefCOCO	SynthRef	DAVIS-2017	J&F↑
✓		val	40.8
✓	✓	val	**44.8**
✓		train+val	33.6
	✓	train+val	27.0
✓	✓	train+val	**38.6**

Table 2: Segmentation accuracy obtained with RefVOS model on two partitionns of DAVIS-2017: validation (*val*) or training+validation (*train+val*). Adding our SynthRef-YouTube-VIS data significantly increases the performance at a zero-cost in annotation. The J&F metric is defined in the supplementary material.

DAVIS-2017 We report the gains of adding the synthetic dataset when evaluating on the standard validation partition of DAVIS-2017 (30 videos), but also on the combined training and validation partitions (90 videos), to obtain more statistically significant results. The results in Table 2 show a significant improvement in segmentation accuracy when adding our synthetic REs to RefCOCO. Figure 2 illustrates qualitative

"a girl in the middle wearing a lab-coat and a black shirt"

"a big man on the right in a black jacket"

(a) Pretrained on RefCOCO   (b) Pretrained on RefCOCO+SynthRef-YouTube-VIS

Figure 2: Qualitative results on DAVIS-2017. Subfigure 2a (left) shows results when the model is pretrained only on RefCOCO, while Subfigure 2b (right) when it is also trained on our synthetic dataset.

Model	SynthRef	J&F↑
(Khoreva et al., 2018)		39.3
(Seo et al., 2020)		44.1
(Bellver et al., 2020)		45.1
(Bellver et al., 2020)	✓	**45.3**

Table 3: Comparison with the state of the art in DAVIS-2017 validation, with models pretrained on RefCOCO and fine-tuned with DAVIS-2017 training data. Adding our generated SynthRef-YouTube-VIS dataset to the RefCOCO pretraining achieves state of the art results. However the relative gain is smaller than in the scenario without fine-tuning, reported in Table 2.

results for this scenario, where the improvement of the segmentation masks is cleary visible. The gain is minor when RefVOS is fine-tuned with training data from DAVIS-2017 and evaluated on the validation partition, as shown in Table 3. This setup is the commonly adopted by the related work, allowing a comparison of our results with them.

RE Source	Prec@0.5↑	Prec@0.9↑	Mean IoU↑
Synthetic	32.3	1.8	35.0
Human	**38.6**	**6.9**	**39.5**

Table 4: Comparison of the performance on a subset of Refer-YouTube-VOS when training with synthetic and human referring expressions.

Refer-YouTube-VOS We further evaluate our method using the subset of Refer-YouTube-VOS that corresponds to our synthetic dataset, SynthRef-YouTube-VIS. We train two instances of RefVOS, one using the human-produced REs of Refer-YouTube-VOS and one using the synthetic REs of SynthRef-YouTube-VIS. The evaluation is done on the test split of SynthRef-YouTube-VIS but using the human REs of Refer-YouTube-VOS

in both models for a fair comparison. Since both human and synthetic expressions (ours) are available for the same videos, we can measure the domain gap between real and synthetic data for training. Our results, reported in Table 4, indicate that, even though the model trained on human referring expressions outperforms the model trained on synthetic ones, the drop in accuracy is not that big to prevent the use of our synthetic data for training. On the contrary, the obtained numbers show that our synthetic expressions can be used interchangeably with the human ones when the latter are hard to acquire.

4 Conclusion

In this work, we propose SynthRef, a novel method for generating synthetic referring expressions, which is used to create the first large-scale dataset of synthetic referring expressions for video object segmentation, namely SynthRef-YouTube-VIS. Our experiments show that pretraining a model using our synthetic referring expressions increases its capability to generalize on new data, which is very important in scenarios where training data are not available for a target dataset. Our method, that does not involve any human annotation cost, can be applied to other existing datasets and tasks (*e.g.* object detection or text-to-image retrieval). We invite the community to explore further possibilities and benefits out of it.

References

Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. 2020. Refvos: A closer look at referring expressions for video object segmentation. *arXiv preprint arXiv:2010.00263*.

Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. 2019. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7454–7463.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. 2013. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199.

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

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. 2018. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1271–1280.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets.

Ehud Reiter and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 1*, COLING '92, pages 232–238, USA. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages –.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725.

C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. 2015. Can humans fly? Action understanding with multiple classes of actors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5188–5197.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10502–10511.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.