SynthRef: Generation of Synthetic Referring Expressions for Object Segmentation (Supplementary Material)

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*

1 SynthRef-YouTube-VIS Dataset

Our proposed method, SynthRef, is applied on YouTube-VIS (Yang et al., 2019), a dataset created from the large-scale video object segmentation dataset YouTube-VOS (Xu et al., 2018). YouTube-VOS is the largest existing benchmark for video object segmentation, including more than 4K high-resolution videos collected from YouTube with a small duration of 3-6 seconds each. Despite containing pixel-level masks for 94 different object categories, YouTube-VOS is not exhaustively annotated, meaning that not all objects belonging to those 94 categories have a corresponding segmentation mask. In contrast, YouTube-VIS, although it has a smaller category set of 40 common objects, it has the advantage that all instances belonging to those categories are labeled. In this way, YouTube-VIS serves as a very good data source for the task of generating synthetic referring expressions, as it is necessary to combine the information of all the present objects in a video frame in order to create valid referring expressions.

YouTube-VIS consists of 2,883 videos with 4,883 unique objects belonging to 40 categories and approximately 131K object masks. The official training set of YouTube-VIS is used for the creation of SynthRef-YouTube-VIS, since ground-truth annotations for all the frames are necessary in order to apply our proposed method. In this way, SynthRef-YouTube-VIS consists of 2,238 videos with 3,374 annotated objects appearing in them. The dataset is further split in a train and test set for the experiments having 1791 training and 447 testing videos. The distribution of object instances of SynthRef-YouTube-VIS over the 40 object classes can be seen in Figure 1. It is observed that all classes appear both in the training and the test set.

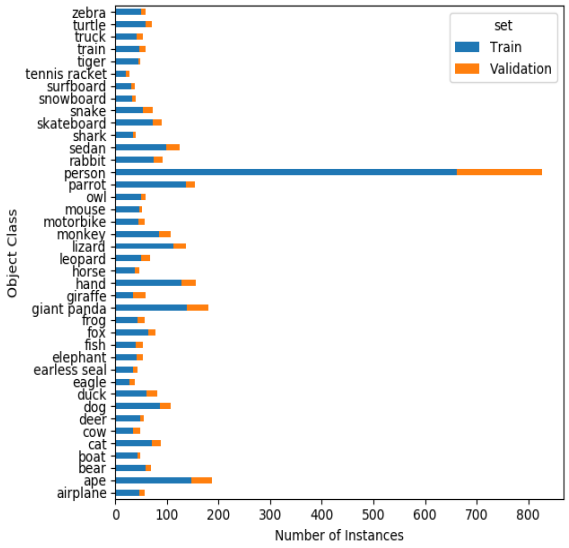

Figure 1: Histogram of object instances by each class in SynthRef-YouTube-VIS train and validation sets.

2 Experiments

2.1 Model

For the evaluation of our method we use RefVOS (Bellver et al., 2020), a frame-based model which uses DeepLabv3 (Chen et al., 2017) with a ResNet-101 (He et al., 2016) backbone as its visual encoder and BERT (Devlin et al., 2019) as its language encoder. The extracted visual and language features are then combined via a multiplication operator and a multi-modal embedding is produced which is fed into a convolutional layer that finally predicts two maps, one for the foreground, which corresponds to the referred object, and another for the background. The ResNet-101 (He et al., 2016) is pretrained on the ImageNet (Deng et al., 2009) dataset. Although the authors of RefVOS report results also freezing the language encoder while training, in our experiments, the language branch is always trained alongside the visual branch.

2.2 Training Details

For training the model, we use a batch size of 8 video frames which are resized and then cropped/padded to a final resolution of 480x480. The optimizer is SGD with a momentum of 0.9 and the learning rate values and schedule depend on the dataset. For training after a first pretraining on RefCOCO, we use an initial learning rate of 1e-4 which is linearly decreased by 4e-6 after every epoch for 20 epochs. For later fine-tuning on DAVIS-2017 the learning rate starts from 1e-5 and is decreased to 1e-6 after 10 epochs. When no pretraining on RefCOCO is performed, an initial learning rate is set at 1e-2 and is decreased by 4e-4 at every epoch for a total of 25 epochs.

2.3 Evaluation metrics

In the task of object segmentation, given a ground truth mask $\mathcal{G}$ and a predicted segmentation mask $\mathcal{M}$, the typical evaluation process includes two measures:

1. Region Similarity $\mathcal{J}$: The similarity of the ground truth and predicted segmentation regions is measured using the Jaccard index $\mathcal{J}$ defined as *Intersection-over-Union (IoU)* of the two regions i.e.:

$$\mathcal{J} = \frac{|\mathcal{M} \cap \mathcal{G}|}{|\mathcal{M} \cup \mathcal{G}|}$$

2. Contour accuracy $\mathcal{F}$: The contour-based precision P_c and recall R_c between the two sets of closed contours $c(\mathcal{M})$ and $c(\mathcal{G})$ for the predicted and ground truth mask respectively, is calculated with an approximation of a bipartite graph matching using morphological operators. Then the typical F-measure (or F_1 score) is defined as:

$$\mathcal{F} = \frac{2P_cR_c}{P_c + R_c}$$

Based on the above measures, the following metrics are being used in our work for the evaluation of our method and the comparison to existing approaches:

- *Precision@X* : Given a threshold *X* in the range [0.5,0.9], a predicted mask for an object is counted as true positive if its $\mathcal{J}$ is larger than *X*, and as false positive otherwise. Then, Precision@X is computed as the ratio between the number of true positives and the total number of instances.
- *Overall* $\mathcal{J}(IoU)$: Total intersection area of all objects divided by the total union area.
- *Mean* $\mathcal{J}(IoU)$: Average of the $\mathcal{J}$ measure (IoU) of all objects so that large and small regions are treated equally.
- $\mathcal{J}\&\mathcal{F}$: The average of the mean region based similarity (Mean $\mathcal{J}$) and the mean contour accuracy (Mean $\mathcal{F}$).

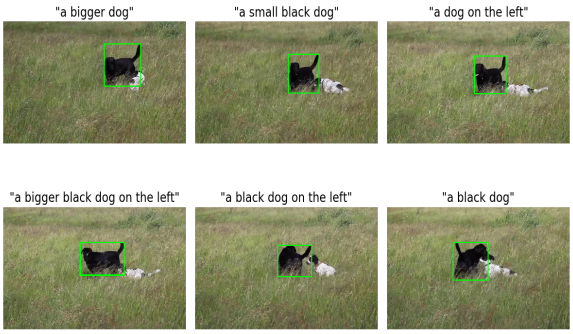

Figure 2: Examples of synthetic referring expressions automatically generated with our method. The target object is indicated with an overlay mask.

3 Synthetic Referring Expressions

An example of different referring expressions generated with our method, for the same video, is illustrated in Figure 1. Multiple referring expressions can be created for the same video or even for the same frame. Moreover, a qualitative comparison of the synthetic referring expression of SynthRef-YouTube-VIS and the human-produced ones of Refer-YouTube-VOS (Seo et al., 2020) for the same videos is illustrated in Figure 3.

4 Referring Expression Information

In order to evaluate the effect of the information included in the synthetic referring expressions, experiments with different amount of information were conducted, starting from a baseline where the synthetic referring expressions consist of just the object class *e.g.* "a dog". Then in the second experiment, relative size and location are added and in the third and last experiment attributes are included too. The model used in these experiments is first pretrained on RefCOCO, then trained with the synthetic referring expressions of SynthRef-YouTube-VIS, using different amount of information in each experiment, as explained

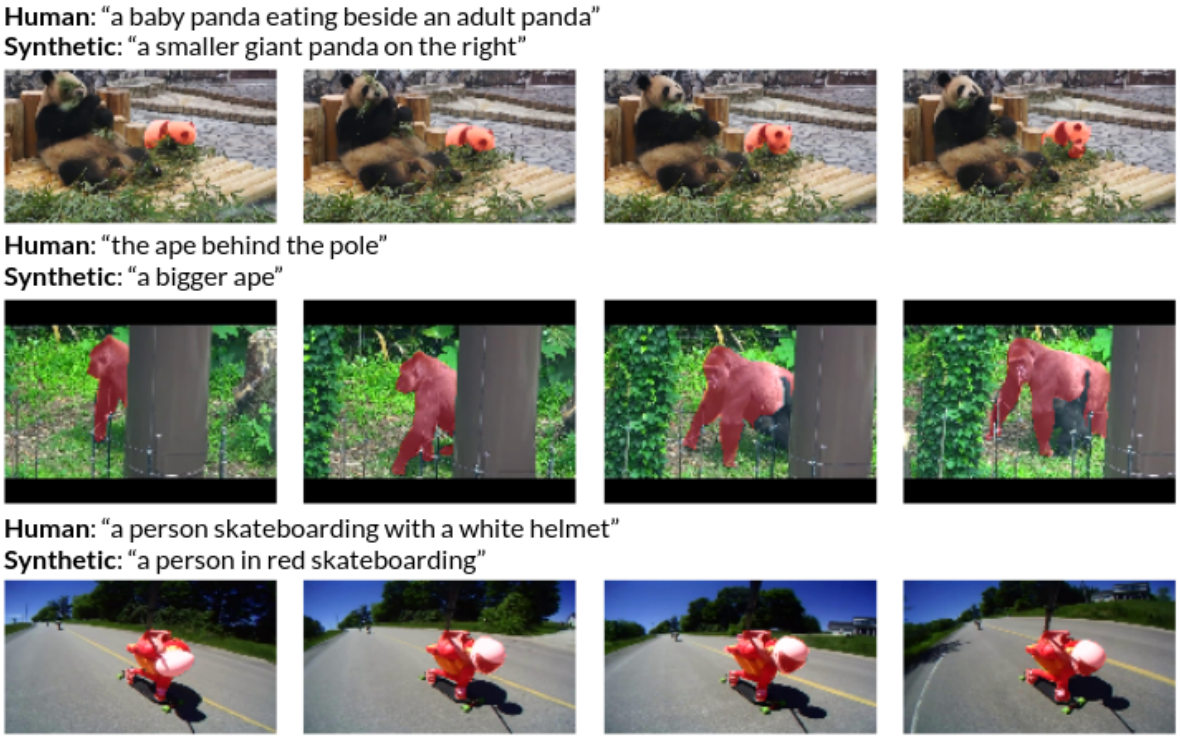

Figure 3: Comparison of human-produced referring expressions of Refer-YouTube-VOS (Seo et al., 2020) with synthetic ones generated with the proposed method.

above, and it is finally fine-tuned on the training set of DAVIS-2017 and evaluated on the validation set.

Referring Expression Information	J&F
Obj. Class	42.0
+ Relative Size + Relative Location	43.5
+ Attributes	45.3

Table 1: Effect of the information included in the synthetic referring expressions on the final performance on DAVIS-2017 validation set.

Results in Table 1 indicate that performance gradually increases with the amount of information provided in the synthetic referring expressions. This is explained because of the fact that in cases where multiple objects of the same class are present in a video, bigger amount of information is necessary in order to unambiguously identify a specific object. It is also remarkable that the final segmentation accuracy in DAVIS-2017 is high, even when only the object class is used as referring expression, during pretraining with SynthRef-YouTube-VIS. This happens for two reasons. The first is that the model is already pretrained on RefCOCO and the second is that DAVIS-2017 validation set includes several videos where only one object instance from each class appears.

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*.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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 –.

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.