SynthRef: Generation of Synthetic Referring Expressions for Object Segmentation


Ioannis Kazakos


Carles Ventura


Miriam Bellver


Carina Silberer


Xavier Giro-i-Nieto

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

National Technical University of Athens

Universitat Oberta de Catalunya

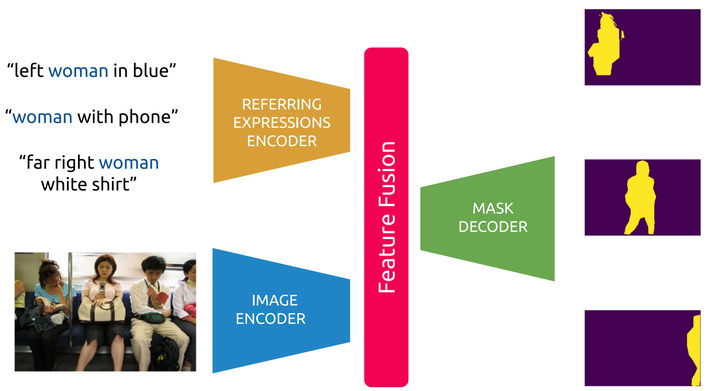
Universität Stuttgart Germany

Barcelona Supercomputing Center Centro Nacional de Supercomputación

Institut de Robòtica i Informàtica Industrial

Task description

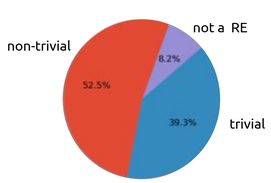
Predict pixel-wise object masks from a **referring expression (RE)**, which disambiguates between instances of a class.



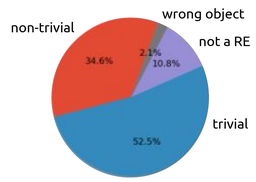
Challenges

REs are costly to obtain.

Existing datasets mostly contain *trivial* expressions [1].



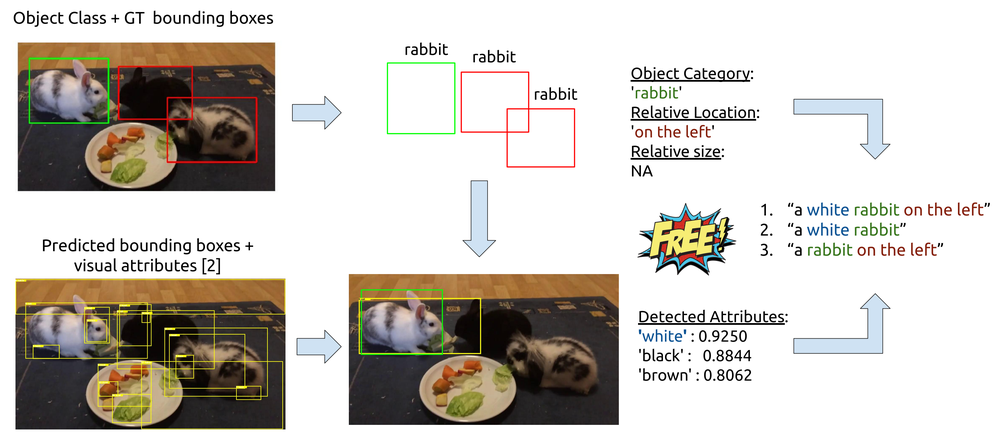
DAVIS 2017 dataset



A2D Sentences dataset

Our approach

Generate **synthetic REs** from the semantic class and bounding boxes already annotated in datasets for large scale object detection.



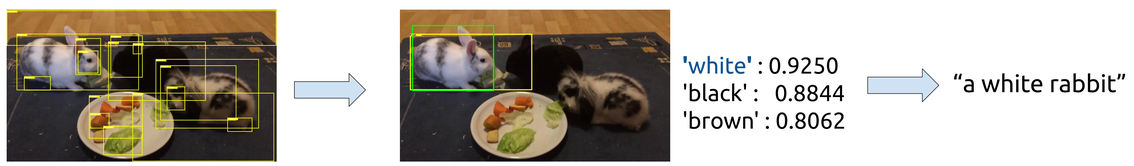
Methodology

SynthRef produces referring expressions based on three cues:

Relative size

Relative location

Discriminative attributes, predicted with [2]



Results

Gain in accuracy in DAVIS-2017 when we add synthetic REs from the YouTube-VIS dataset to pre-train the RefVOS model [1] for video object segmentation.

We measure the **domain gap** when training RefVOS with synthetic (free) or human generated (costly) REs from Refer-YouTube-VOS [3].

Accuracy on DAVIS-2017 train+val

Pretraining	J&F
RefCOCO	33.6
SynthRef-YouTube-VIS	27.0
RefCOCO+SynthRef-YouTube-VIS	**38.6**

Accuracy on Refer-YouTube-VOS

Ref. Expressions	Prec@0.5	Prec@0.9	Mean IoU
Synthetic	32.27	1.82	35.02
Human	**38.61**	**6.87**	**39.46**

[1] Bellver, M., Ventura, C., Silberer, C., Kazakos, I., Torres, J., Giro-i-Nieto, X. **RefVOS: A Closer Look at Referring Expressions for Video Object Segmentation**. arXiv 2020.
[2] Tang, K., Niu, Y., Huang, J., Shi, J., & Zhang, . **Unbiased scene graph generation from biased training**. CVPR 2020.
[3] Seo, Seonguk, Joon-Young Lee, and Bohyung Han. **"URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark."** ECCV 2020.

Erasmus+ NVIDIA GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD Generalitat de Catalunya