Chosen Final Project :

Topic J - STCN Video Segmentation

Group members :

Lacombe Yoach
Venard Paul-Louis

Plan of work :

The plan of work, i.e. what are you going to implement, what data you are going to use, what experiments you are going to do. If working in a group, who are the members of the group and how you plan to share the work. The project proposal will represent 10% of the final project grade.

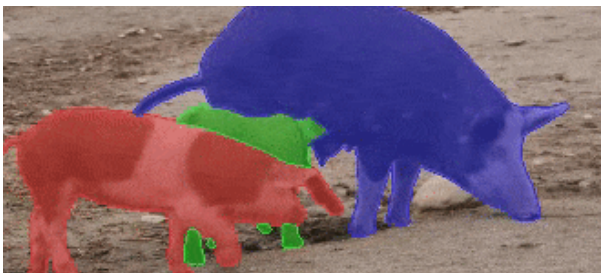
As stated in the description, we will first use the authors' implementation [5], trying to replicate their results and verify the performance of their model on another couple video dataset/corresponding bounding box annotation like provided in [6][7].

Then, we will implement state-of-the-art pretrained models for semantic segmentation like the famous Mask-RCNN[2], or try new approaches like PoinRend Head[3] and Swin-Transformer NLP-like Backbone[4] to automatise the initialization of the video segmentation method. The obtained results will be compared to existing one in [1] and those obtained in the first part.

Regarding the share of the workload, as the first part seems easier and shorter than the second one, we are expecting that one of us will work on the first part and will provide help and support to the second one that will exclusively work on the second part.

Reminder of the Project :

Motivation



The visual segmentation of the scene has become crucial for many applications including autonomous driving or learning from demonstrations. One of the state-of-the-art video segmentation methods [1] assumes that the initial frame annotation is known. This initialization is used to segment the whole video sequence. This algorithm has been shown to be robust and fast but requires manual one-frame annotation. Your goal will be to propose an automatization of the first frame annotation via state-of-the-art image segmentation methods, *e.g.* [2,3,4].

Description

The first goal of the project is to replicate some of the results from the paper at the inference time, *i.e.* to select one of the tested datasets and verify authors' results by using authors' implementation [5]. The second objective of the project is to verify performance on additional

video dataset that was not used in [1], *e.g.*, the something-something dataset [6] and the corresponding bounding box annotations for computing approximate qualitative results [7]. The final objective is to use one of the existing image segmentation approaches for the initialization of the video segmentation method and compare it to manual annotation results.

References

[1] Cheng, Ho Kei, Yu-Wing Tai, and Chi-Keung Tang. "Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation." arXiv preprint arXiv:2106.05210 (2021).

[2] https://github.com/matterport/Mask_RCNN

[3] https://github.com/facebookresearch/detectron2/tree/main/projects/PointRend

[4] https://github.com/SwinTransformer/Swin-Transformer-Semantic-Segmentation

[5] https://hkchengrex.github.io/STCN/

[6] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The “something something” video database for learning and evaluating visual common sense. In ICCV, 2017.

[7] J. Materzynska, T. Xiao, R. Herzig, and H. Xu. Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks. In CVPR, 2020. https://github.com/joaanna/something_else