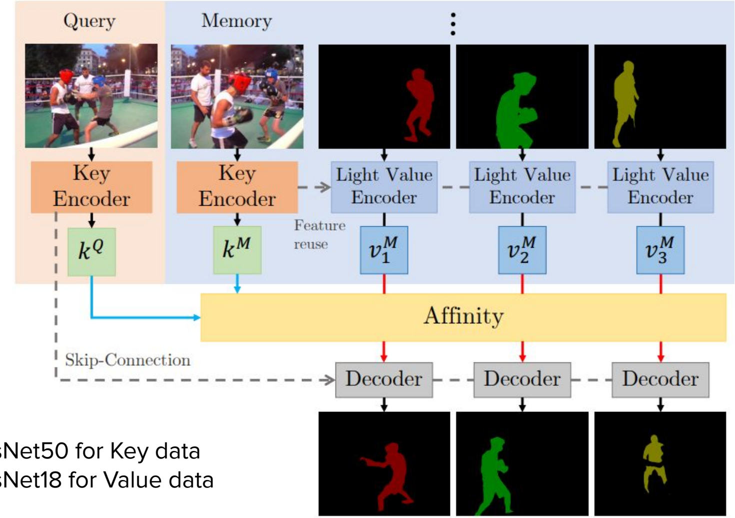
STCN video segmentation

Angelika ANDO - angelika.ando@ens-paris-saclay.fr
Yujin CHO - yujin.cho@ens-paris-saclay.fr

école
normale
supérieure
paris—saclay

Space-Time Correspondence Networks (STCN)



ResNet50 for Key data
ResNet18 for Value data

STCN architecture . Extract from “Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation” by Cheng et al.

- Simpler, more efficient, faster than STM
- Negative squared Euclidean distance as a similarity measure
- Robust Affinity
- Less Memory usage than STM
- Used authors’ implementation

<u>Study goals :</u>

- Verify author’s results
- Study reaction of STCN to new dataset
- Measure result impact when model is feeded by different techniques
    - Mask generated by segmentation algorithms

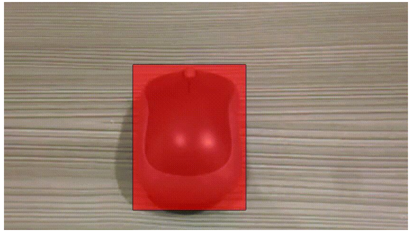
Task 1 - Verify authors' implementation

Method	DAVIS 2017			
	$\mathcal{J}\,\&\,\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
Original paper	**85.3**	**82.0**	**88.6**	**20.2**
Our experiments	**85.3**	**82.0**	**88.6**	**16.9**

- FPS differences can be due to cloud environment
- GPU efficiency for processing time

Task 2 - Verify performance on additional dataset using bounding-boxes

- New dataset : Something-Else (13 videos)
- Only Bounding-box available as ground truth
- Correct result for objects with simple shape (eg.: mouse)
- STCN succeed to track roughly the object and adapt masks



example of good behavior

example of bad behavior

Task 3 - First frame initialization by segmentation algorithm

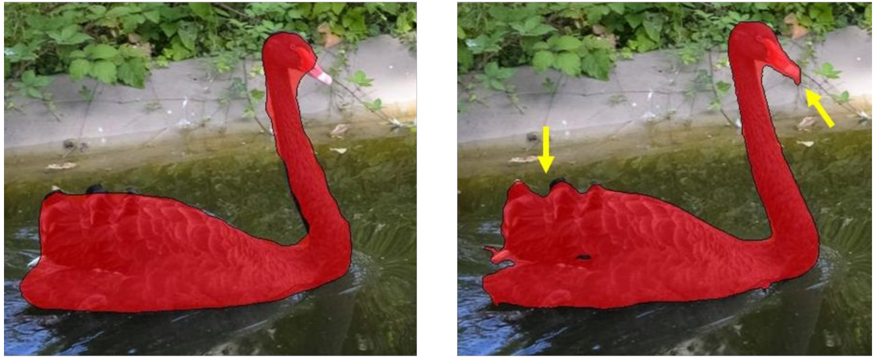
Mask-R-CNN (2018) (evolution of Faster-R-CNN)

- Backbone ResNet50 pretrained on ImageNet
- Model pretrained on COCO train 2017
- Output : Boxes / Scores / Labels / Masks
- Mask resolution 28 x 28
- Pre-processing and Segmentation algorithm self written code

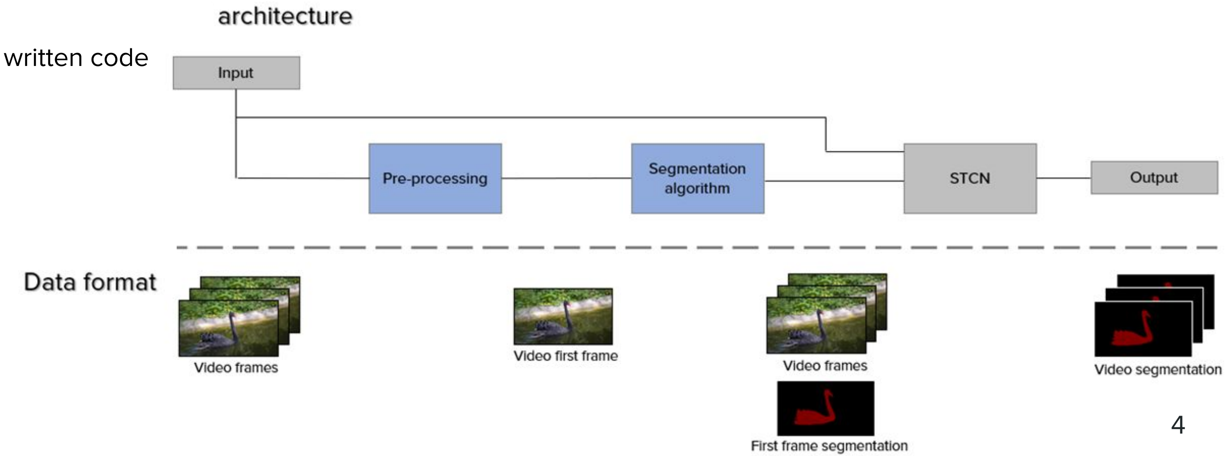
PointRend (2020) (evolution of Mask-R-CNN)

- Point Head module over a Mask-R-CNN pretrained on COCO
- Output : Boxes / Scores / Labels / Masks
- Mask resolution 224 x 224
- Pre-processing and Segmentation algorithm self written code

Segmentation example on Davis2017 dataset



Mask-R-CNN output PointRend output



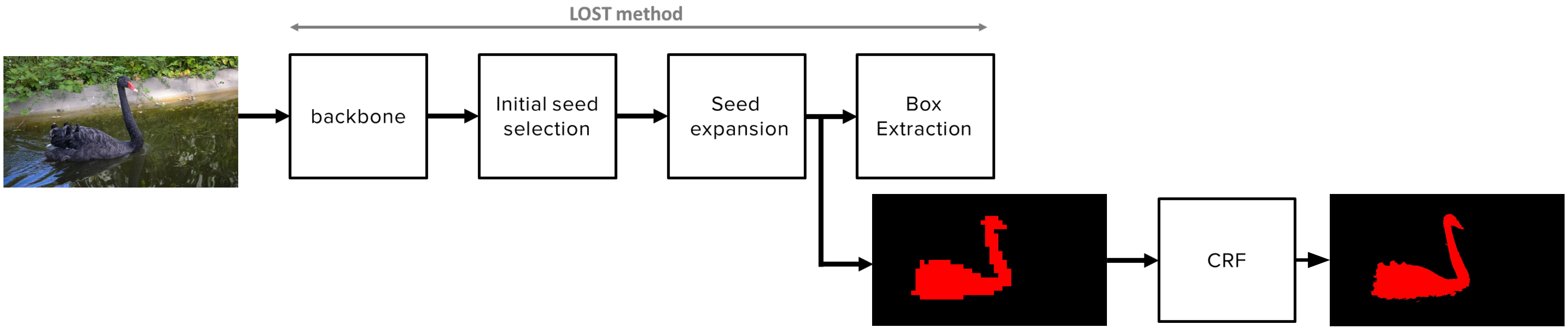
Task 3 - First frame initialization by segmentation algorithm

LOST (Localizing Objects with Self-Supervised Transformers and no Labels, 2021)

- Unsupervised learning method
- Backbone VIT-S/16 trained with DINO method
- Output : Boxes / Scores / Labels / Masks
- Assumption : a patch with low correlation belongs to an object than to the background
- Adapted authors' implementation

CRF (Conditional Random Fields, 2012)

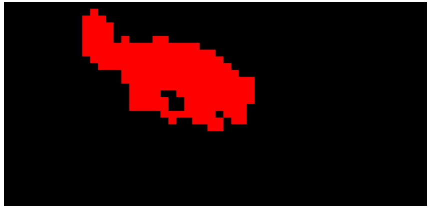
- Undirected probabilistic graphical model
- Models P(Y|X) : X - observation (pixel colour), Y - segmentation label per pixel
- Adapted authors' implementation



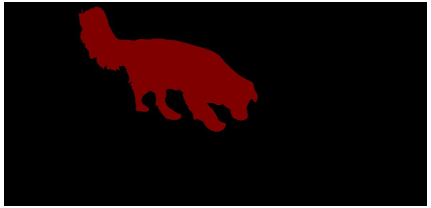
Task 3 - First frame initialization by segmentation algorithm

Example mask obtained by LOST + CRF

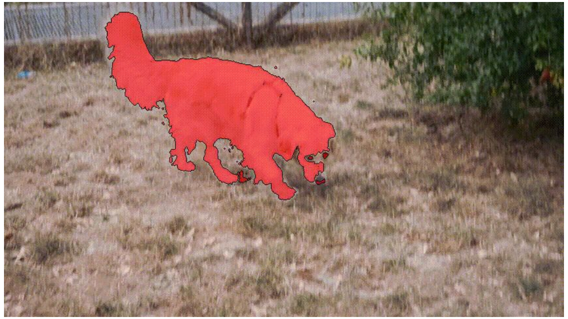

Image


LOST


LOST + CRF


Original mask

Example videos obtained by LOST + CRF



$\mathcal{J}$	$\mathcal{F}$
95.7	96.3

$\mathcal{J}$	$\mathcal{F}$
95.4	96.9

$\mathcal{J}$	$\mathcal{F}$
0.001	0.037

Task 3 - Results

- Better results compared to Task 2 (supposing that we do not have ground truth segmentation for the first frame)

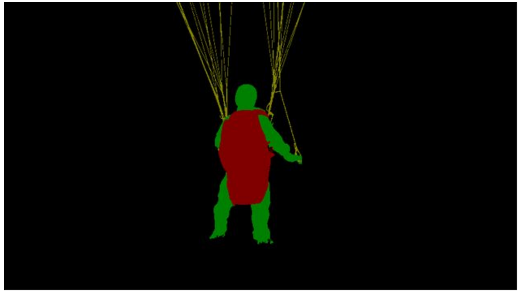
- Comparing Mask R-CNN, PointRend and LOST methods

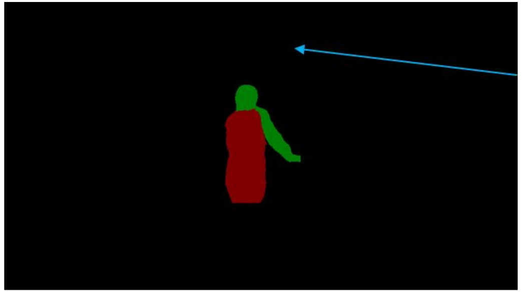
Method	DAVIS 2017			
	$\mathcal{J}\,\&\,\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
Ground truth 1st frame	**85.3**	**82.0**	**88.6**	16.9
Mask R-CNN 1st frame	69.9 ↓	67.9 ↓	72.0 ↓	17 ↑
Pointrend 1st frame	71.1 ↓	68.6 ↓	73.6 ↓	14.9 ↓
LOST + CRF 1st frame	25.7 ↓	23.9 ↓	27.6 ↓	**19.9** ↑

PointRend results

Task 3 - Drawbacks on segmentation methods

- Impossibility for segmentation algorithm to find « unusual » objects

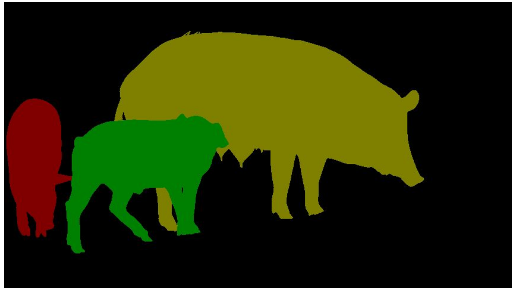

1st frame ground truth mask

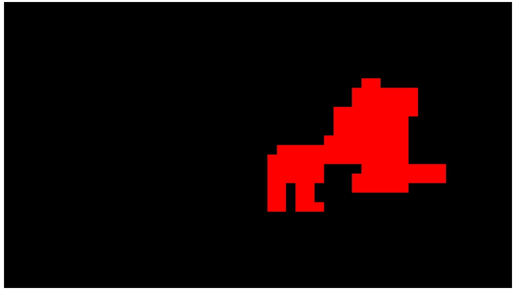

Mask generated by PointRend

Missing parachute

Partial/small objects or unknown labels are more difficult for segmentation algorithm to be found

- LOST has trouble with multiple similar objects


1st frame ground truth mask


Mask generated by LOST

LOST works well if there is one object

When there are more objects, LOST usually detects only one, or make a mask containing more objects

If there is a large object in the image, LOST algorithm tends to confuse it with the background as the feature patches will have a bigger correlation rate.

Conclusion

- Difficulty to manage big datasets
- Satisfying visual results
- Many code adaptation
- STCN is robust even with many objects to track
- STCN performs well even if the first frame annotation is not perfect
- STCN is state of the art SVOS model

STCN result with many objects to track

References

PAPER

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)

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.

J. Materzynska, T. Xiao, R. Herzig, and H. Xu. **“Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks”**. In CVPR, 2020.

Oriane Simeoni and Gilles Puy and Huy V. Vo and Simon Roburin and Spyros Gidaris and Andrei Bursuc and Patrick Perez and Renaud Marlet and Jean Ponce, **"Localizing Objects with Self-Supervised Transformers and no Labels"** In BMVC, 2021

GITHUB

https://hkchengrex.github.io/STCN

https://github.com/davisvideochallenge/davis2017-evaluation

https://github.com/joaanna/something_else

https://github.com/matterport/Mask_RCNN

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

https://github.com/valeoai/LOST

https://github.com/lucasb-eyer/pydensecrf