STCN Video Segmentation
Final Project Report

26 January 2022

Angelika Ando
ENS Paris-Saclay
angelika.ando@ens-paris-saclay.fr

Yujin Cho
ENS Paris-Saclay
yujin.cho@ens-paris-saclay.fr

Abstract

A video segmentation task is to identify a set of objects in a video scene and to track their movements. In this work, we present the STCN video segmentation method which is part of SVOS (Semi-Automatic Video Object Segmentation). Deep learning-based SVOS methods use the first frame annotation mask to segment the whole video sequence. The STCN model shows state of the art performance, and our goal is to present various experiments on the STCN method.

1. Introduction

STCN (Space Time Correspondence Network) [9] is simpler, more efficient and faster than STM (Space-Time Network) [15], which was introduced in 2019. Figure 1 shows the architecture of STCN. Each new video frame, also named query, passes through a key encoder to construct an affinity matrix between the query and the memory keys. This affinity matrix, based on the negative L_2 distance, captures the correspondences between objects. STCN uses only one affinity matrix to gain memory and computation efficiency unlike STM. For encoders, STCN has a lighter network than STM. At the final stage of decoding, STCN generates the new masks. Also, it does not need the last frame key and values unlike STM and only depends on the affinity matrix. Thus, it allows to have a better memory efficiency.

This work is divided into 3 parts. First, we verify the authors' implementation [1] on the DAVIS 2017 dataset. Second, we study the performance of STCN on 13 videos of the Something-Something dataset [10] using the corresponding bounding-box annotations of the Something-Else dataset [14]. The 13 videos and the bounding-box annotations were given in [2]. Third, we investigate results when the first frame annotation is given by different segmentation algorithms, namely Mask R-CNN [3, 11], Pointrend [4, 12] and LOST [5, 16] combined with the CRF [6, 13] processing step. For the third part, we compare these algorithms both for the DAVIS 2017 dataset using segmentation masks as first frame annotations and for the Something-Something dataset using bounding-boxes as first frame annotations.

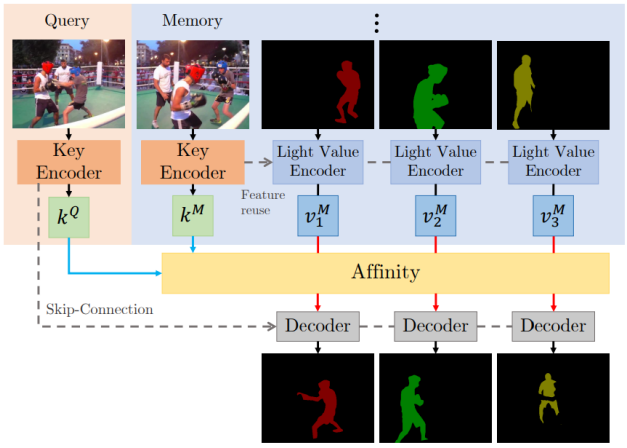


Figure 1: Space Time Correspondence Networks (STCN) architecture. Extracted from [9] by Cheng et al.

2. Related works

2.1. Mask R-CNN

Mask R-CNN is an extension of Faster R-CNN and has a branch for predicting segmentation masks on each Region of Interest (RoI). It uses ResNet50 as a backbone. Faster R-CNN only predicts class and bounding box coordinates because it is not designed for pixel to pixel alignment between input and output. To overcome this problem, Mask R-CNN proposes the RoI Align module that can preserve the exact spatial locations and predict a mask for each object found.

2.2. PointRend

PointRend works well both for instance segmentation (e.g., Mask R-CNN) and semantic segmentation (e.g., FCN). The backbone of PointRend produces a coarse mask and the PointRend module refines it by upsampling it by 2 using linear interpolation. Point sampling selects a non-uniform set of points that has a high probability to change the label value and it computes the new label. This process is repeated until reaching the final resolution wished (224×224). Therefore, detailed la-

beling is possible and it increases the classification accuracy of the border.

2.3. LOST

LOST is an unsupervised learning method that is based on the assumption that a patch with low correlation in the image is more likely to belong to an object than to the background. Moreover, two patches within the same object are likely to correlate positively with each other, and correlate negatively with the background.

The first step of LOST is to apply a vision transformer to the image and to extract the keys $k_1, \ldots, k_N$ of the last attention layer, where N is the number of patches. These keys are the feature vectors of the patches and LOST uses these features to calculate the correlation (dot product) between the patches. Then LOST follows in three steps: 1) seed selection, 2) seed expansion and 3) mask and bounding-box extraction.

Seed selection consists in generating the patch similarity graph. There is an edge between two patches if their features are positively correlated. The seed will be a patch with lowest degree. **Seed expansion** is the step when we choose the next best seeds. A patch is chosen as seed if it is within the $k = 100$ patches with lowest degree and it is positively correlated with the initially chosen seed. Finally, we select further patches which are positively correlated with the mean feature of the selected seeds. Then we select the largest connected component containing the first seed to remove fake patches. The remaining patches constitute the **mask** that we further refined with CRF but they are also used to **extract the bounding-boxes**.

2.4. CRF

CRFs (Conditional Random Fields) are undirected probabilistic graphical models that take the context into account. The CRF processing step used in this project models $\mathbb{P}(Y|X)$, where X is the observed pixel colours of the initial mask and Y is the segmentation label per pixel for the new mask.

The maximum a posteriori segmentation label is calculated by minimizing the Gibbs energy. Gibbs energy consists of two parts: 1) unary potential at each pixel, which is based on the initial segmentation mask and 2) pairwise potentials between pairs of pixels. The pairwise potentials penalise small segmentation regions and they force close pixels of similar colours to have the same class.

3. Implementations

Code is available here : `https://github.com/nuniniyujin/STCN_evaluation`

3.1. Mask R-CNN and Pointrend

In order to compare the video segmentation between the original implementation (giving ground truth mask as first frame) and first frame mask generated by a segmentation model, some adaptations were needed. We implemented the architecture presented in Figure 2 using `pytorch` library. The **input** is composed of all RGB frames of the video, one mask per object and per frame, and one object bounding box for each frame.

In the **Pre-processing** stage we will only extract the first frame RGB info and first frame object bounding box info. The RGB image will be denormalized (Denormalized Image $=$ Image $\times$ std $+$ mean) as the segmentation models require each channel value to be between $(0, \ldots, 255)$.

In the **Segmentation algorithm** block, we call a pre-trained segmentation model and we give the RGB image as input. We decided to do the segmentation on the whole image as it corresponds more to a realistic approach. We are not supposed to have access to the bounding box of the object in real situation. In our study, we filter the output of the segmentation result by calculating the L_2 distance between the ground truth bounding box and the one found by the segmentation model. The set of first frame masks is then computed to be the input of STCN. This method was used for both Mask R-CNN and PointRend and the result can be seen in Figure 3, where a big number of output objects were filtered to keep only the object needed to calculate accuracy.

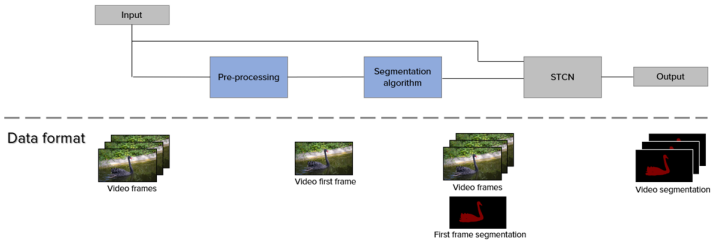


Figure 2: Self-supervised architecture using segmentation model to produce an object mask for the first frame.

Figure 3: Results of filtering Mask R-CNN outputs.
(From left to right : before filtering Mask R-CNN outputs, after filtering outputs, ground truth mask.)

3.2. LOST and CRF

We adapted the authors' implementation of LOST [5] and CRF [6]. The implementation details follow the procedures described in Sections 2.3 and 2.4. For the Something-Something dataset we used the bounding-boxes predicted by LOST. For the DAVIS 2017 dataset we extracted the mask which is produced after seed expansion and which LOST uses for the bounding box predictions afterwards. Then we further refined this mask with CRF. Figure 4 shows the procedure described above.

For extracting the features of the patches, LOST uses a ViT-S model pre-trained with DINO [8]. It extracts the keys of the last attention layer and these keys are the feature vectors of the patches. We tried to extract the queries and the values of the last

attention layer to compare them with the keys. Interestingly, the queries and the values give worse resulting masks than the keys.

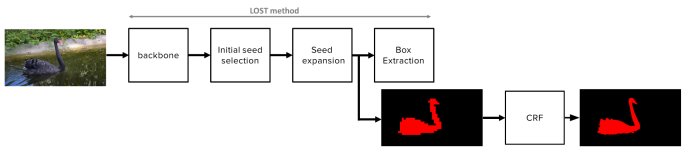


Figure 4: Mask and box extraction using LOST and CRF.

Figure 5: Comparing the original mask to the mask obtained by LOST and CRF.

4. Experiments

In our experiments, we use the same evaluation metrics as in paper [9] for the task to verify paper performance.

- Region Jaccard $\mathcal{J}$: It is calculated by intersection over Union (IoU) between segmentation results and ground truth. It is called region similarity as well.
- Boundary Accuracy $\mathcal{F}$: Harmonic mean of boundary precision P_c and recall R_c. It represents how well the segment contours match with the ground-truth contours.
- $\mathcal{J}$ & $\mathcal{F}$: Both $\mathcal{J}$ and $\mathcal{F}$ are used for evaluation.
- FPS : Number of frames per second. The larger it is the smoother and better the displaying quality is.

Video results are available at : `https://bit.ly/32mezrQ`

4.1. Reproduction of results

For reproducing the results of STCN, we used the evaluation code found in [7]. Our experiment results correspond to the original paper results. The only difference is FPS which can be explained by different hardware used to compute.

1st frame	$\mathcal{J}$ & $\mathcal{F}$ ↑	$\mathcal{J}$ ↑	$\mathcal{F}$ ↑	FPS ↑
Original paper	85.3	82.0	88.6	20.2
Our experiments	85.3	82.0	88.6	16.9

Table 1: Results of STCN on the DAVIS 2017 dataset using manual first frame annotations.

4.2. Results on the DAVIS 2017 Dataset

Table 2 shows the results of different segmentation algorithms. PointRend has a slightly better result than Mask R-CNN in terms of $\mathcal{J}$ & $\mathcal{F}$ thanks to the high-resolution mask output. Using segmentation algorithms for the first frames has some drawbacks. Sometimes it fails to find small or partial objects, which leads to lower values of $\mathcal{J}$ & $\mathcal{F}$.

LOST combined with CRF performed very well when there was only one object in the image. In contrast, it had difficulties in detecting multiple objects as it either predicted a mask containing more objects or it only detected a small part of one of them. The low quantitative results are due to these phenomena. More examples are provided in the Appendix.

1st frame	$\mathcal{J}$ & $\mathcal{F}$ ↑	$\mathcal{J}$ ↑	$\mathcal{F}$ ↑	FPS ↑
Ground truth	85.3	82.0	88.6	16.9
Mask R-CNN	69.9	67.9	72	17
Pointrend	71.1	68.6	73.6	14.9
LOST + CRF	25.7	23.9	27.6	19.9

Table 2: Results of STCN on the DAVIS 2017 dataset using segmentation masks as first frame annotations.

4.3. Results on the Something-Something Dataset

Table 3 shows the results of different segmentation algorithms applied to extract bounding-box annotations for the first video frames. These low results can be easily explained. The ground truth of this dataset is only bounding boxes. When we provide STCN a bounding box, sometimes the model will adapt it to the shape of the object (see appendix B.1.), which will lead to a good video segmentation but a poor accuracy for boundary scores as we compare the segmentation with the bounding boxes of the objects. Another problematic point is the lack of depth information: we cannot know which object is in front of the other. We may have overlapped object masks and make the model mistaken when extracting object features.

1st frame	$\mathcal{J}$ & $\mathcal{F}$ ↑	$\mathcal{J}$ ↑	$\mathcal{F}$ ↑	FPS ↑
Ground truth	41.4	53.0	29.8	19.3
Mask R-CNN	29.2	39.7	18.8	25.8
Pointrend	27.0	35.1	18.9	23.4
LOST	17.2	20.9	13.5	28.8

Table 3: Results of STCN on the Something-Something dataset using bounding-boxes as first frame annotations.

5. Conclusion

In this project, we studied the STCN algorithm, which allows to segment whole video sequences given only the annotation of the first video frame.

Most of our experiments where done on the DAVIS 2017 dataset. We compared the performance of STCN using manual and automatized segmentation masks for the first frame. We found that PointRend worked better than Mask R-CNN as it has a better mask output resolution. The quality of the first frame mask has the main impact on the performance of STCN as it is the main input for the computation of the affinity matrix. LOST combined with CRF worked well for segmenting one object but had difficulties with multiple or large objects.

We also compared the performance of STCN on 13 videos of the Something-Something dataset using bounding boxes as first

frame annotations. For simple shaped objects, STCN succeeded to adapt the bounding box mask to the shape of the object, but with multiples objects STCN reacted poorly.

We can conclude that STCN is robust even when there are many objects to track and it performed well even if the first frame annotation was not perfect.

References

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

[2] https://github.com/joaanna/something_else.

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

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

[5] https://github.com/valeoai/LOST.

[6] https://github.com/lucasb-eyer/pydensecrf.

[7] https://github.com/davisvideochallenge/davis2017-evaluation.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[9] H. K. Cheng, Y.-W. Tai, and C.-K. Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021.

[10] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

[13] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24:109–117, 2011.

[14] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[15] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[16] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Appendix

Video results are available at: `https://bit.ly/32mezrQ`

A. Examples of Qualitative Results for the DAVIS 2017 Dataset

A.1. Comparing the mask obtained by Mask R-CNN and PointRend to the original mask.

	$\mathcal{J}$	$\mathcal{F}$
Original	96.6	99.4
Mask R-CNN	91.3	94.8
PointRend	94.6	98.2

Figure 6: Segmentation example on a single object: PointRend mask is sharper and has high frequency variations.

	$\mathcal{J}$	$\mathcal{F}$
Original	93.7	97.3
Mask R-CNN	87.3	92.7
PointRend	90.2	94.2

Figure 7: Segmentation example on multiple objects: PointRend mask is sharper and has high frequency variations.

	$\mathcal{J}$	$\mathcal{F}$
Original	88.1	90.4
Mask R-CNN	75.7	76.3
PointRend	49.5	62.9

Figure 8: Segmentation example where two objects are interacting.

	$\mathcal{J}$	$\mathcal{F}$
Original	44.5	51.1
Mask R-CNN	21.7	25.8
PointRend	38.2	39.4

Figure 9: Segmentation example with blurry image.

A.2. Comparing the mask obtained by LOST and CRF to the original mask.

	$\mathcal{J}$	$\mathcal{F}$
Original	96.0	98.5
LOST + CRF	95.4	96.9

Figure 10: Good mask prediction for the dog.

	$\mathcal{J}$	$\mathcal{F}$
Original	97.1	99.4
LOST + CRF	95.1	96.6

Figure 11: Good mask prediction for the camel.

	$\mathcal{J}$	$\mathcal{F}$
Original	96.7	99.7
LOST + CRF	94.5	95.3

Figure 12: CRF managed to remove the hiker's shadow.

	$\mathcal{J}$	$\mathcal{F}$
Original	96.6	99.7
LOST + CRF	0.00	0.00

Figure 13: One of the small lamps was selected as initial seed as it had the lowest correlation in the image. After taking the largest connected component containing the seed in the patch similarity graph, the resulting mask is only a patch. CRF further refined this patch to one pixel.

	$\mathcal{J}$	$\mathcal{F}$
Original	93.7	97.3
LOST + CRF	0.10	3.70

Figure 14: Only half of one pig was detected. CRF is good in refining masks but not in extending them.

	$\mathcal{J}$	$\mathcal{F}$
Original	88.1	90.4
LOST + CRF	47.7	63.6

Figure 15: A mask covering both of the people.

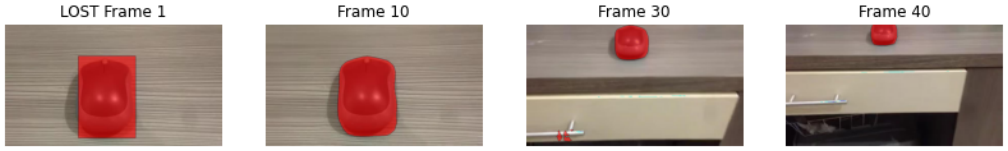
	$\mathcal{J}$	$\mathcal{F}$
Original	84.0	93.8
LOST + CRF	0.10	0.80

Figure 16: Detection of a wrong object.

B. Examples of Qualitative Results for the Something-Something Dataset

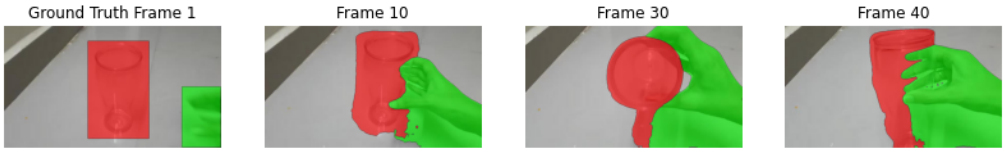
B.1. Comparing the results of STCN using different first frame annotations.

	$\mathcal{J}$	$\mathcal{F}$
Original	83.0	56.0

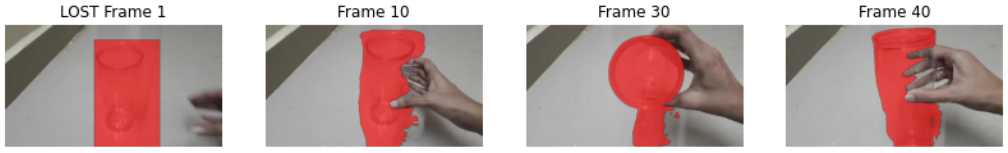


	$\mathcal{J}$	$\mathcal{F}$
LOST	82.6	52.7

Figure 17: The mouse is tracked well in all cases.



	$\mathcal{J}$	$\mathcal{F}$
Original	42.1	18.3



	$\mathcal{J}$	$\mathcal{F}$
LOST	36.9	7.80

Figure 18: The mask of the glass is adapted by STCN but the tracking of the hand is not perfect. LOST only detected one object.

	$\mathcal{J}$	$\mathcal{F}$
Original	59.9	35.1

	$\mathcal{J}$	$\mathcal{F}$
LOST	63.4	27.0

Figure 19: STCN did not adapt the mask for the hammer.