ICCV 2021, Structured Representations for Video Understanding Workshop

SFTrack++: A Fast Learnable Spectral Segmentation Approach for Space-Time Consistent Tracking

Elena Burceanu
eburceanu@bitdefender.com

Bitdefender,
Romania

University of Bucharest,
Romania

Motivation and contributions

Tracking by taking advantage of the **intrinsic object consistency** over **space and time**

- Using SFSeg*, a fast 3D spectral segmentation method over the video's graph of pixels

Improving the 3D spectral formulation for segmentation

- Enable **multiple inputs** towards **robustness**
- Enable **learning**, by make the algorithm differentiable

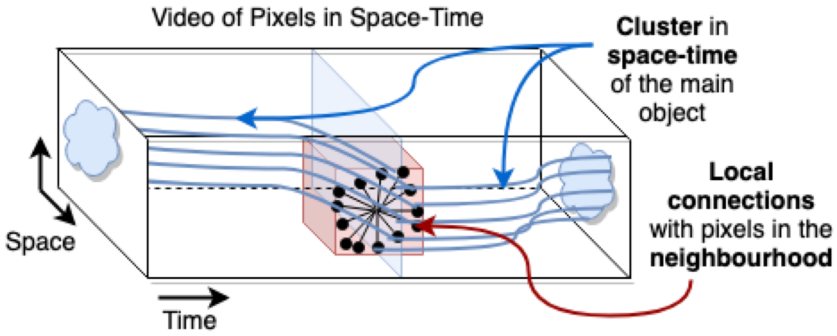
Challenge the **rough bounding boxes** used for tracking

- Learn end-to-end, going through segmentation as an intermediate representation

*Learning a Fast 3D Convolutional Approach to Spectral Object Segmentation in Space and Time, E. Burceanu, M. Leordeanu, 2020

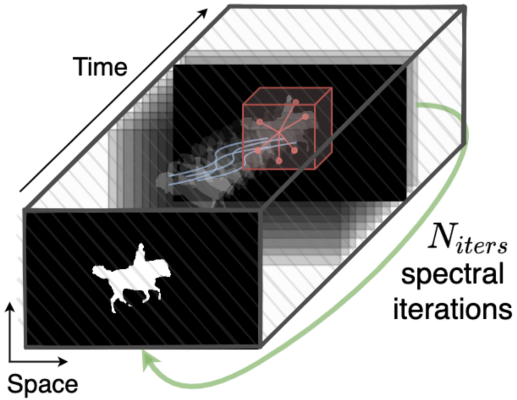
Prior work on spectral segmentation*

- See the video as a volume of pixels
  - **Interconnected** in space and time
  - With an intrinsic **graph structure**



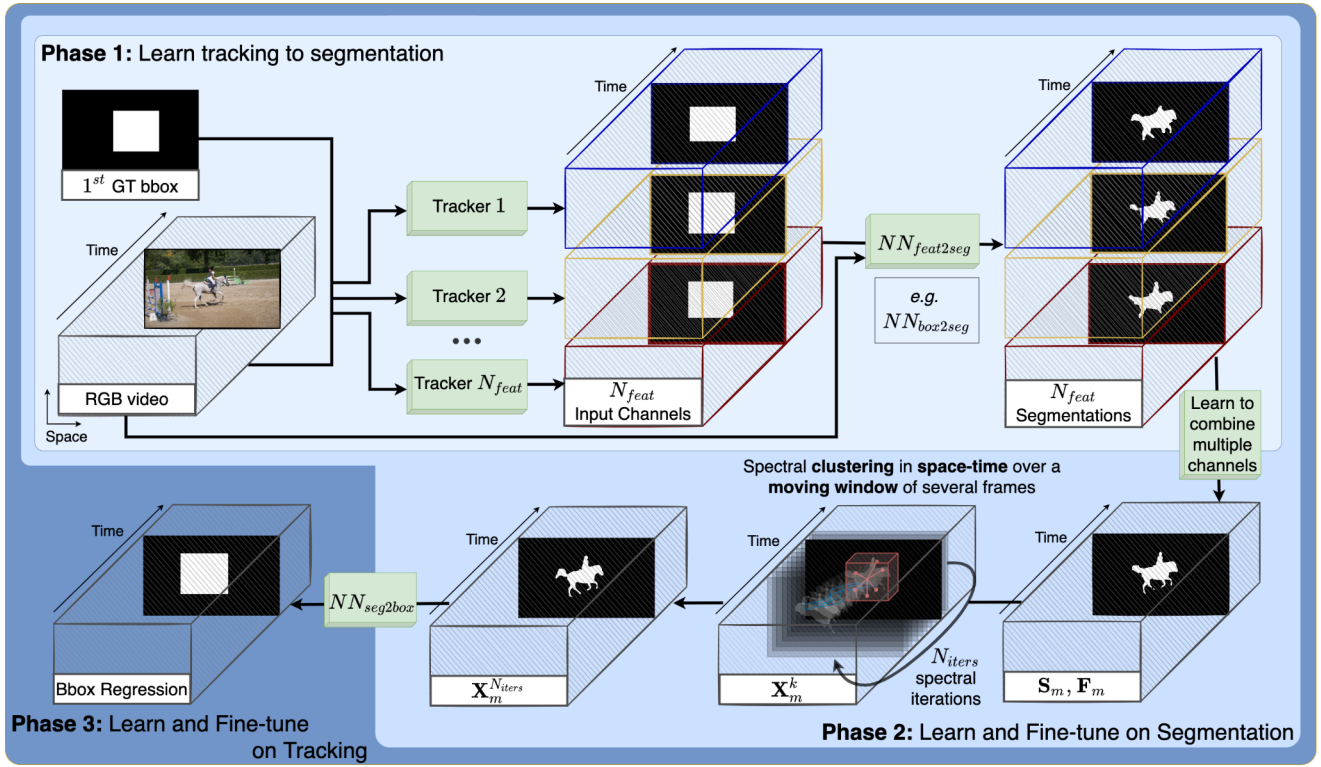
Object segmentation in video

- The **strongest cluster** in the volume
- Strongest cluster is given by the **leading eigenvector**



*Learning a Fast 3D Convolutional Approach to Spectral Object Segmentation in Space and Time, E. Burceanu, M. Leordeanu, 2020

Approach


Phase 1: Learn tracking to segmentation
1st GT bbox
Time
RGB video
Space
Tracker 1
Tracker 2
Tracker N_{feat}
N_{feat} Input Channels
$NN_{feat2seg}$
e.g. $NN_{box2seg}$
N_{feat} Segmentations
Learn to combine multiple channels
Spectral clustering in space-time over a moving window of several frames
N_{iters} spectral iterations
$\mathbf{S}_m, \mathbf{F}_m$
$\mathbf{X}_m^k$
$\mathbf{X}_m^{N_{iters}}$
$NN_{seg2box}$
Bbox Regression
Phase 3: Learn and Fine-tune on Tracking
Phase 2: Learn and Fine-tune on Segmentation

Algorithm

Phase 1. Segmentation

- use multiple trackers as input
- learn $NN_{feat2seg}$: RGB + tracking features or output $\rightarrow$ segmentation mask

Phase 2. Segmentation

- learn to combine multi-channels for spectral space-time consistency
- fine-tune all
- apply fast power iteration over a moving window of frames

Phase 3. Tracking

- learn $NN_{seg2box}$
- fine-tune end-to-end

Results

- For **tracking**, train only on GOT-10k and TrackingNet training sets (5 frames/video)
- **Robust (low variance) state-of-the-art results** on **GOT-10k** and **TrackingNet**
- **UAV**: datasets with a **different data distribution** (smaller objects, captured from drones)

	Method	OTB	UAV	NFS	GOT-10k			TrackingNet		
		AUC	AUC	AUC	AO	SR_{50}	SR_{75}	Prec	$Prec_{norm}$	AUC
Single Method	D3S	57.7	45.0	38.6	39.3	39.0	10.1	52.2	67.9	52.4
	SiamBAN	67.6	60.8	54.2	54.6	64.6	40.5	68.4	79.5	72.0
	ATOM-18	66.7	64.3	58.4	55.0	62.6	39.6	64.8	77.1	70.3
	SiamRPN++	65.0	**65.0**	50.0	51.7	61.5	32.5	**69.3**	80.0	73.0
	PrDimp-18	**67.6**	63.5	**62.6**	**60.8**	**71.0**	**50.3**	69.1	**80.3**	**75.0**
Ensemble	Basic (median)	66.6	60.8	55.5	54.7	63.9	31.6	69.0	80.0	73.9
	Neural Net	**71.3**	59.7	58.2	59.5	69.8	42.9	70.6	80.2	74.5
	SFTrack++	70.3	**61.2**	**62.4**	**62.0**	**73.3**	**47.8**	**71.9**	**81.9**	**76.1**
	std	± 0.5	± 0.2	± 0.1	± 0.7	± 0.5	± 1.1	± 0.3	± 0.3	± 1.0

Ablations

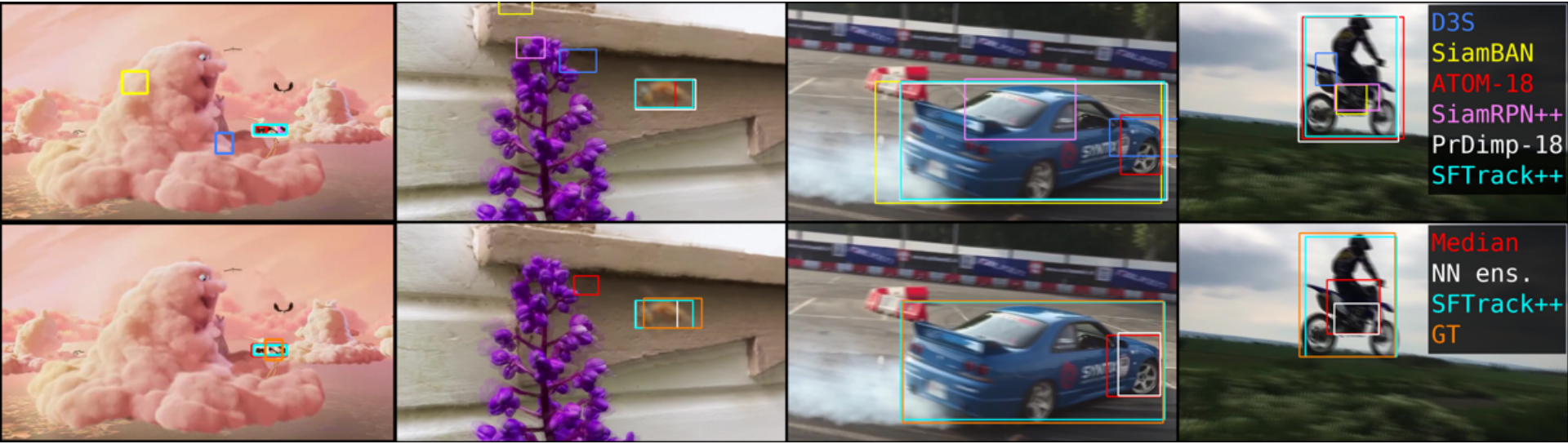
- Remove **pipeline components**
- **Input methods** combinations
- Vary the **number of spectral iterations**

SFTrack++ variations	**OTB**	**UAV**	**NFS**	**OTB+UAV+NFS**
w/o Spectral Refinement (phase two)	**71.6**	60.5	60.9	64.0
w/o $\text{NN}_{segm2bbox}$ (phase three)	65.5	57.4	58.5	60.2
Median (over 5 methods) as input	70.8	60.8	60.0	63.7
Best method (PrDimp-18) as input	67.1	59.7	61.3	62.5
Top 3 methods as input	64.8	60.8	61.1	62.1
2 spectral iterations	70.3	60.9	61.8	64.1
3 spectral iterations	68.0	61.0	60.1	62.9
SFTrack++ (1 iter, 5 methods)	70.3	**61.2**	**62.4**	**64.5**

Qualitative results

SFTrack++ combines the input methods better

- even when there is a **high variance** among them.



SFTrack++ - Takeaway message

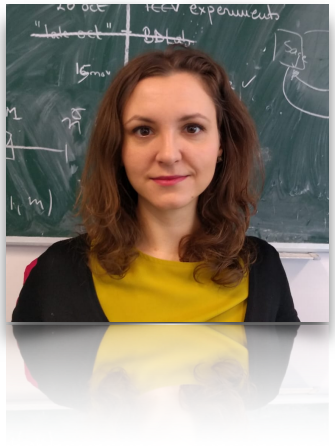
- Preserve the tracked **object consistency over space and time**
  - Adapt 3D spectral filtering algorithm to
    - **combine multiple** and powerful inputs
    - to allow an end-to-end **learning pipeline**

- Use segmentation as an **intermediate representation** for the tracking task
  - **Improved** performance
  - **Robust** solution

Thank you!

Code: https://github.com/bit-ml/sftrackpp

If you have any **questions** or just want to reach out:
https://ilarele.github.io



Elena Burceanu
eburceanu@bitdefender.com

Bitdefender,
Romania

University of Bucharest,
Romania