

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

Elena Burceanu

EBURCEANU@BITDEFENDER.COM

Bitdefender

University of Bucharest, Romania

Institute of Mathematics of the Romanian Academy

Abstract

We propose an object tracking method, SFTrack++, that smoothly learns to preserve the tracked object consistency over space and time dimensions by taking a spectral clustering approach over the graph of pixels from the video, using a fast 3D filtering formulation for finding the principal eigenvector of this graph's adjacency matrix. To better capture complex aspects of the tracked object, we enrich our formulation to multi-channel inputs, which permit different points of view for the same input. The channel inputs are in our experiments, the output of multiple tracking methods. After combining them, instead of relying only on hidden layers representations to predict a good tracking bounding box, we explicitly learn an intermediate, more refined one, namely the segmentation map of the tracked object. This prevents the rough common bounding box approach to introduce noise and distractors in the learning process. We test our method, SFTrack++, on five tracking benchmarks: OTB, UAV, NFS, GOT-10k, and TrackingNet, using five top trackers as input. Our experimental results validate the pre-registered hypothesis. We obtain consistent and robust results, competitive on the three traditional benchmarks (OTB, UAV, NFS) and significantly on top of others (by over 1.1% on accuracy) on GOT-10k and TrackingNet, which are newer, larger, and more varied datasets.

1. Introduction

Better using the temporal aspect of videos in visual tasks has been actively discussed for a rather long time, especially with the large and continuous progress in hardware. The first aspect we tackle in our approach is a seamless blending of space and time dimensions in visual object tracking. Current methods mostly rely on target appearance and frame-by-frame processing [Caelles et al. \(2017\)](#); [Li et al. \(2019\)](#); [Danelljan et al. \(2019\)](#), with rather few taking explicit care of temporal consistency [Bhat et al. \(2020\)](#); [Jabri et al. \(2020b\)](#). In the spectral graph approach, nodes are pixels and edges are their local relations in space and time, while the strongest cluster in this graph, given by the principal eigenvector of the graph's adjacency matrix, represents the consistent main object volume over space and time.

A second observation challenges the rough bounding box (bbox) shape used for tracking. While it provides a handy way to annotate datasets, it is a rather imperfect label since it leads to errors that accumulate, propagate, and are amplified over time. Objects rarely look like boxes, and bboxes contain most of the time significant background information or

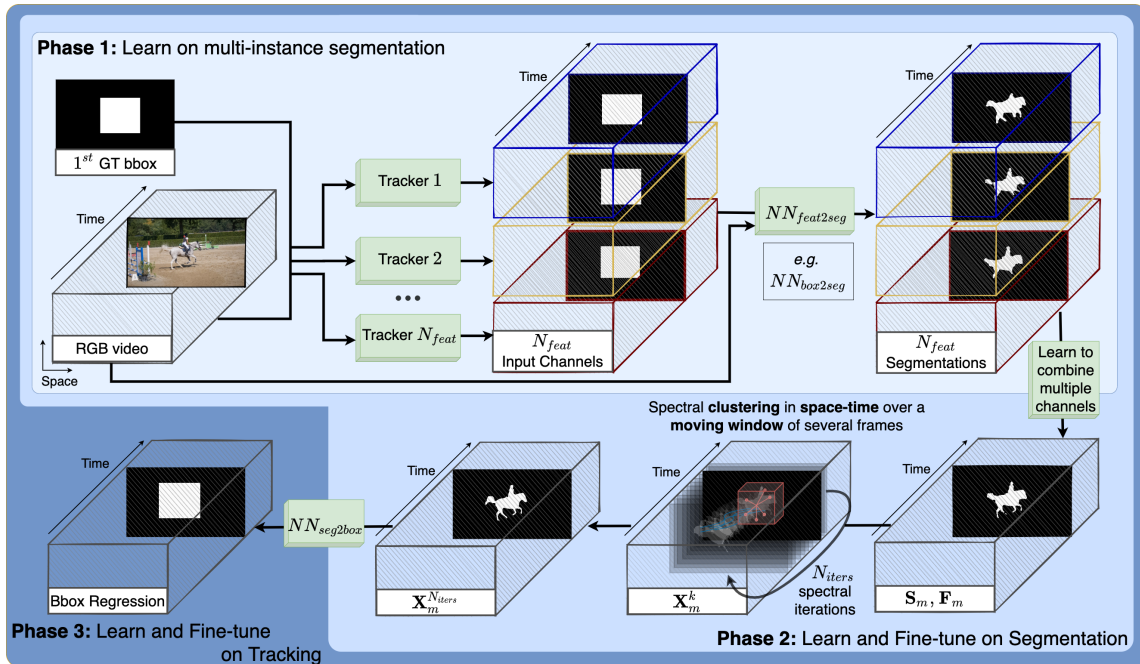


Figure 1: SFTrack++: We start from video’s RGB and 1st frame GT bbox of the tracked object. We run state-of-the-art trackers, in an online manner, while fine-tuning $NN_{feat2seg}$ network frame-by-frame (pretrained in Phase 1) to transform the extracted feature maps (e.g. bboxes) to segmentation maps. Next, we learn to combine multiple segmentation inputs and refine the final mask using a spectral approach, applied also online over a moving window containing the previous N frames, for N_{iter} spectral iterations (Phase 2). In Phase 3, we learn a bbox regressor from the final segmentation mask, $NN_{seg2box}$ and fine-tune all our parameters on the tracking task.

distractors. Since having a good segmentation for the interest object directly influences the tracking performance, we constrain an intermediary representation, a segmentation map, which aims to reduce the quantity of noise transferred from a frame to the next one. We integrate it into our end-to-end flow, as shown in Fig. 1.

A third point we emphasize on is relying on multiple, independent characteristics of the same object or multiple modules specialized in different aspects. This comes with an improved ability to understand complex objects while increasing the robustness [Bhat et al. \(2020\)](#). We, therefore, adjust the SFSeg spectral approach, enhancing its formulation to support learning on top of multiple input channels. They could be general features for the input frame coming from different approaches or, more specifically, tracking outputs from multiple solutions, as we test in Sec. 4.

The **main contributions** of our approach are:

- SFTrack++ brings to tracking a natural, contiguous, and efficient approach for integrating space and time components, using a fast 3D spectral clustering method over the graph of pixels from the video, to strengthen the tracked object’s model.
- We explicitly learn intermediate fine-grained segmentation as opposed to rough bounding boxes in our three phases end-to-end approach to a more robust tracking solution.
- We integrate into our formulation a way of learning to combine multiple input channels, offering a wider view of the objects, harmonizing different perceptions, for a powerful and robust approach.

2. Related work

General Object Tracking. Out of the three main trackers families, **Siamese based trackers** gained a lot of traction in recent years for their high speed and end-to-end capabilities Bertinetto et al. (2016b); Li et al. (2018, 2019); Zhang and Peng (2020). Most approaches focus on exhaustive offline-training, failing to monitor changes w.r.t. the initial template Tao et al. (2016); Xu et al. (2020); Chen et al. (2020), while others update their model online Voigtlaender et al. (2020); Dai et al. (2020). Nevertheless, the robustness towards unseen objects and transformations at training time remains a fundamental problem for Siamese trackers. **Meta-learning approaches for tracking** Park and Berg (2018); Bertinetto et al. (2016a); Wang et al. (2020) come with an interesting way of adapting to the current object of interest, while keeping a short inference time, by proposing a target-independent tracking model. One major limitation for both those approaches, Siamese and meta-learning trackers, is that they fail to adapt continuously to the real-time changes in the tracked object, using rather a history of several well-chosen patches or even just the initial one. In contrast, our method naturally integrates the temporal dimension, by continuously enforcing the local temporal and spatial object consistency. **Discriminative methods** Danelljan et al. (2019, 2020); Lukezic et al. (2020) on the other hand are classic approaches, focusing more on changes in the tracked object Danelljan et al. (2017) (background, distractors, hard negatives), better integrating the temporal dimension Bhat et al. (2020) in the method flow. They prove to be robust, but they mostly rely on hand-crafted observations or modules not trainable end-to-end. SFTrack++ provides an end-to-end approach while minimizing the distractors and background noise using the intermediary segmentation map. Our method distances itself from a certain family of trackers, introducing the space and time consistency endorsement via clustering component as an additional dimension of the algorithm.

With a few notable exceptions Wang et al. (2019); Voigtlaender et al. (2019), most tracking solutions use internally hidden layer representations extracted from the previous frame’s rough bbox prediction Bhat et al. (2020); Danelljan et al. (2020, 2019); Xu et al. (2020), rather than a fine-grained segmentation mask as in our approach. Also, most of them do not take into account multiple perceptions for the input frame and operate over a unique feature extractor Danelljan et al. (2019); Zhang and Peng (2020); Bertinetto et al. (2016b). There are a few trackers though that combine two models for adapting to sudden changes while remaining robust to background noise, by explicitly model the different pathways Bhat

et al. (2020); Burceanu and Leordeanu (2018). In contrast, our end-to-end multi-channel formulation learns over 10 input channels, a significantly larger number.

Graph representations. Images and videos were previously represented as graphs, where the nodes are pixels, super-pixels, or regions Jabri et al. (2020a). This choice directly impacts the running time and performance. Regarding edges, they are usually undirected, modeled by symmetric similarity functions, but there are also several works that use directed ones Torsello et al. (2006); Yu and Shi (2001). **Spectral clustering** approaches Ng et al. (2001); Meila and Shi (2001); Leordeanu and Hebert (2005) search for the leading or the smallest eigenvector for the graph’s adjacency matrix to solve the clusters’ assignments. Spectral clustering was previously used in pixel-level image segmentation Shi and Malik (2000), with a high burden on the running time and in building space-time correspondences between video patches Jabri et al. (2020a). Graph Cuts is a common approach for spectral clustering, having many variations Shi and Malik (2000); Ding et al. (2001); Sarkar and Soundararajan (2000). SFSeg Burceanu and Leordeanu (2020) proposes a 3D filtering technique for efficiently finding the spectral clustering solution without explicitly computing the graph’s adjacency matrix. Inspired by this method, we integrate an improved version with learning over multi-channel inputs, as an intermediate component in our tracker, as detailed in Sec. 3.

3. Our approach

SFTrack++ algorithm has three phases, as we visually present them in Fig. 1. In **Phase 1**, we learn a neural net, $NN_{feat2seg}$, that transforms the RGB and a frame-level feature map extracted using a tracker (*e.g.* bbox from a tracker prediction) into a segmentation mask. Using only the RGB as input is not enough, because frames can contain multiple objects and instances, and we also need a pointer to the tracked object to predict its segmentation. Next, in **Phase 2**, we run multiple state-of-the-art trackers frame-by-frame over the input as an online process and extract input channels from them (*e.g.* bboxes). We transform those feature maps to segmentation maps with the previously recalled module, $NN_{feat2seg}$. Next, we learn to combine and refine the outputs for the current frame using a spectral solution for preserving space-time consistency, adapted to learn over multiple channels. Note that, when applying the spectral iterations, we use a sliding window approach over the previous N frames in the video volume. For supervising this path, we use segmentation ground-truth. **Phase 3** learns a neural net as a bbox regressor over the final segmentation map from the previous phase, $NN_{seg2box}$, while fine-tuning all the other trainable parameters in the model, using tracking GT.

Spectral approach to segmentation. We go next through the following aspects, briefly explaining the connection between them: segmentation \rightarrow leading eigenvector \rightarrow power iteration \rightarrow 3D filtering formulation \rightarrow multi-channel. Image segmentation was previously formulated as a graph partitioning problem, where the segmentation solution Shi and Malik (2000) is the leading eigenvector of the adjacency matrix. It was used in a similar way for video Burceanu and Leordeanu (2020). Power iteration algorithm can compute the leading eigenvector: $\mathbf{x}_i^{k+1} \leftarrow \sum_{j \in \mathcal{N}(i)} \mathbf{M}_{i,j} \mathbf{x}_j^k$, where \mathbf{M} is the $N \times N$ graph’s adjacency matrix, N is the number of nodes in the graph (pixels in the video space-time volume in our case), $\mathcal{N}(i)$ is the space-time neighbourhood of node i and each step k is followed by normalization.

The adjacency matrix used in power iteration usually depends on two types of terms: unary ones are about individual node properties and pairwise ones describe relations between two nodes (pairs).

Following this approach, SFSeg rewrites power iteration using 3D filtering for an approximated adjacency matrix. The solution is described in Eq. 1:

$$\mathbf{X}^{k+1} \leftarrow \text{normalized}(\mathbf{S}^p \cdot (\alpha^{-1}\mathbf{1} - \mathbf{F}^2) \cdot G_{3D} * (\mathbf{S}^p \cdot \mathbf{X}^k) - \mathbf{S}^p \cdot G_{3D} * (\mathbf{F}^2 \cdot \mathbf{S}^p \cdot \mathbf{X}^k) + 2\mathbf{S}^p \cdot \mathbf{F} \cdot G_{3D} * (\mathbf{F} \cdot \mathbf{S}^p \cdot \mathbf{X}^k)), \quad (1)$$

where $*$ is a 3D convolution with Gaussian filter G_{3D} over space-time volume, \cdot is an element-wise multiplication, \mathbf{S} and \mathbf{F} are unary and pairwise terms in matrix form with p and α controlling their importance, k is the current spectral iteration and $\mathbf{X}, \mathbf{S}, \mathbf{F}$ matrices have the original video shape ($N_{frames} \times H \times W$).

Multi-channel learning formulation. We extend the single-channel formulation in SF-Seg such that it can learn how to combine several input channels, S_i and F_i , for unary and pairwise terms respectively: $\mathbf{S}_m \leftarrow \sigma(\sum_{i=1}^{N_{cs}} w_{s,i} \mathbf{S}_i + b_s \mathbf{1})$, $\mathbf{F}_m \leftarrow \sigma(\sum_{i=1}^{N_{cf}} w_{f,i} \mathbf{F}_i + b_f \mathbf{1})$, where \mathbf{S}_m and \mathbf{F}_m are the multi-channel unary and pairwise maps, respectively, σ is the sigmoid function, N_{cs} and N_{cf} are the number of input channels, $\mathbf{1}$ is an all-one matrix for the bias terms and $w_{s,i}, w_{f,i}, b_s, b_f$ are their corresponding learnable weights. We replace \mathbf{S} and \mathbf{F} in Eq. 1 with their multi-channel versions \mathbf{S}_m and \mathbf{F}_m , respectively. We learn $w_{s,i}, w_{f,i}, b_s, b_f$ parameters both over segmentation and tracking tasks. More, SFTrack++ can learn end-to-end, from the original input frames all the way to final output, in the case of end-to-end learnable feature extractors.

4. Experimental protocol

We test if SFTrack++ brings in a complementary dimension to tracking by having an intermediary fine-grained representation, extracted over multiple state-of-the-art trackers’ outputs, and smoothed in space and time. We guide our experiments such that we evaluate the least expensive pathways first. For reducing the hyper-parameters search burden, we use AdamW Loshchilov and Hutter (2019), with a scheduler policy that reduces the learning rate on a plateau. For efficiency and compactness, we use the same channels to construct both the unary and pairwise maps: $\mathbf{S}_i = \mathbf{F}_i$. Their learned weights are also shared $w_{s,i} = w_{f,i}$. We use as input channels bboxes extracted with top single object trackers. We choose 10 top trackers: SiamR-CNN Voigtlaender et al. (2020), LTMU Dai et al. (2020), KYS Bhat et al. (2020), PrDiMP Danelljan et al. (2020), ATOM Danelljan et al. (2019), Ocean Zhang and Peng (2020), D3S Lukezic et al. (2020), SiamFC++ Xu et al. (2020), SiamRPN++ Li et al. (2019), SiamBAN Chen et al. (2020), which differ in architecture, training sets and overall in their approaches, but all achieves top results on tracking benchmarks.

Training. In Phase 1 we train our $NN_{feat2seg}$ network on DAVIS-2017 Pont-Tuset et al. (2017) and Youtube-VIS Yang et al. (2019) trainsets, for each individual object. It receives the current RGB and the output of a tracking method (bbox), randomly sampled at training time. We use the U-Net architecture, validating the right number of parameters (100K - 1 mil) and the number of layers. We use DAVIS-2017 and Youtube-VIS evaluation sets to stop the training. Following the curriculum learning approach, before introducing tracking methods into the pipeline, we use at the beginning of the tracking GT bboxes (extracted

from segmentation GT, as straight bboxes). This allows the $NN_{feat2seg}$ component to get a good initialization, before introducing faulty bbox extractors, namely the top 10 tracking methods mentioned before. For Phase 2, we also train for the segmentation task. We learn the second part of our method to have an intermediary fine-grained representation, extracted over multiple channels, and smoothed in space and time. We validate here N_{iters} , the number of spectral iterations (1-5). We train on DAVIS-2016 [Perazzi et al. \(2016\)](#) and Youtube-VIS datasets. In Phase 3, training for tracking, we learn a regression network, $NN_{seg2bbox}$ (with 50K-500K parameters), to transform the final segmentation to bbox. We train on TrackingNet [Müller et al. \(2018\)](#), LaSOT [Fan et al. \(2019\)](#) and GOT-10k [Huang et al. \(2019\)](#) training splits.

Baselines Comparison. Experiment for comparing with other methods focus on the improvements SFTrack++ could bring over state-of-the-art and other competitive approaches for general object tracking: single method state-of-the-art solutions, a basic ensemble over the trackers, SFTrack++ applied only over the best tracker, SFTrack++ applied over the basic ensemble and the best learned neural net ensemble we could get out of several configurations (2D and 3D versions for U-Net [Ronneberger et al. \(2015\)](#) and shallow nets, having a different number of parameters: 100K, 500K, 1 mil, 5 mil, 15 mil). All methods receive the same input from top 10 trackers and train on TrackingNet, LaSOT and GOT-10k train sets as previously described. We evaluate our solution against all baselines on seven tracking benchmarks: **VOT2018** [Kristan et al. \(2018\)](#), **LaSOT**, **TrackingNet**, **GOT-10k**, **NFS** [Galoogahi et al. \(2017\)](#), **OTB-100** [Wu et al. \(2015\)](#) and **UAV123** [Mueller et al. \(2016\)](#). For the main conclusion of the paper, we will provide statistical results (mean and variance over several runs) to better indicate a strong positive/negative result, or an inconclusive one.

Ablative studies. We vary several components of our end-to-end model to better understand their role and power. We train our **Phase 1** component, $NN_{feat2seg}$ net, not only for bbox input features but also for other earlier features, extracted from each tracker architecture. We test the overall tracking performance for this case. We remove from the pipeline the spectral refinement in **Phase 2** and report the results. We test the performance of our tracker without the **Phase 3** neural net, $NN_{seg2bbox}$, by replacing it with a straight box and rotated box extractors from OpenCV [Bradski and Kaehler \(2008\)](#). We test several losses to optimize for both segmentation and tracking tasks: a linear combination between the weighted dice loss [Sudre et al. \(2017\)](#) and binary cross-entropy, Focal-Tversky [Abraham and Khan \(2019\)](#), and Focal-Dice [Wang and Chung \(2018\)](#). For the ablative experiments, we evaluate only on OTB100, UAV123, and NFS30 tracking datasets.

5. Experimental Results

Comparison with other methods. In Tab. 1 we present the results of our method on five tracking benchmarks: OTB100, UAV123, NFS30, GOT-10k, and TrackingNet, comparing it with other top single methods and ensemble solutions. For single methods, we take into account each input method in our SFTrack++: D3S, SiamBAN, ATOM-18, SiamRPN++, PrDimp-18. We chose only one lightweight configuration per tracker, common across all benchmarks. For ensembles we use 1) a basic per-pixel median ensemble followed by the same bounding box regressor used in all experiments (see Sec. 7) and we also trained a

	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

Table 1: Comparison on 5 tracking benchmarks. In the first group we show individual methods, used as input for our SFTrack++. In the second one, we show ensemble methods: a basic (median) and a neural net model with a similar number of parameters like SFTrack++. Our method outperforms both the input or other ensemble methods by a large margin on the challenging benchmarks GOT-10k and TrackingNet, while obtaining competitive results on OTB, UAV, and NFS. For SFTrack++ we report mean and std when training the model from scratch three times. With blue we represent the best single method in the column and with red the best ensemble. The raw results are available in the supplementary material.

more complex one: 2) a neural net having an UNet architecture (with 5 down-scaling and 5 up-scaling layers) and a similar number of parameters like SFTrack++ (≈ 4.3 millions). We observe that the variation across different runs (including training from scratch) of our method is very small, showing a robust result and a clear conclusion. On newer, larger, and more generic datasets like GOT-10k and TrackingNet, our method surpasses others by a large margin, while on OTB100, UAV123, and NFS30 it has competitive results.

Ablation studies. To validate the components of our method, we test in Tab. 2 different variations, reporting results on OTB100, UAV123, and NFS30. First, we remove the spectral refining component from phase 2, taking out the temporal dependency and leaving the per frame predictions independent. In the next experiment, we remove the neural net from phase3 $NN_{segm2bbox}$. In the next chunk, we investigate the number of input methods. Last, we vary the number of spectral iterations from phase 2. The conclusions from this wide ablation are the following: 1) the spectral refining component is very important, emphasizing the initial intuition that preserving the object consistency in space and time using our proposed spectral approach improves the overall performance in tracking. 2) The quality of the input in our SFTrack++ method is important, but the more methods we use, the better. 3) We obtain better results using only one single spectral iteration. We tried

SFTrack++ variations	OTB	UAV	NFS	OTB+UAV+NFS
w/o Spectral Refinement (phase two)	71.6	60.5	60.9	64.0
w/o 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

Table 2: Ablations on OTB100+UAV123+NFS30 benchmarks. In the first group we remove from the pipeline phase 2 and phase 3, respectively. The results show that both components are crucial for the method. Next, we vary the number of input methods (1 to 5) but also their quality (best or median). We see that even the quality of the input matters, using more methods as input improves the overall performance. In the last part, we validate the number of spectral iterations, a single iteration achieving the best score, which slowly degrades over more iterations.

the loss functions mentioned in the protocol, but we did not see relevant variations in IoU for the validation set, so we settle to BCE.

Qualitative results. Since SFTrack++ is an ensemble method, we show in Fig. 2 difficult cases and how it compares with individual methods and with other ensembles, starting from the same input (namely, all single methods from the first line, as explained in Sec. 5). We see how our method outperforms the others, even in those hard cases where an agreement seems hard to achieve.

6. Findings

The pre-registered hypotheses that SFTrack++, our spectral approach that improves the space-time consistency of an object, improves the tracking performance proved to be valid.

Our method is not choosing the best input method, but it learns how to combine all inputs towards a superior performance, not only w.r.t. each input, but also w.r.t. a basic and a learned ensemble solution. SFTrack++ is robust, with a very low variance when re-training the entire pipeline from scratch, as shown in Tab. 1. We also noticed that the output of our method is very bold and it does not depend on a carefully chosen threshold. In Tab. 2 we show the importance of integrating the spectral clustering module in our SFTrack++ algorithm.

As a side observation, SFTrack++ has a weak performance on very small tracked objects when compared with single methods (UAV videos), but when compared with other ensembles, it achieves top results. This might be due to the large variance in the chosen individual methods predictions for the small objects.

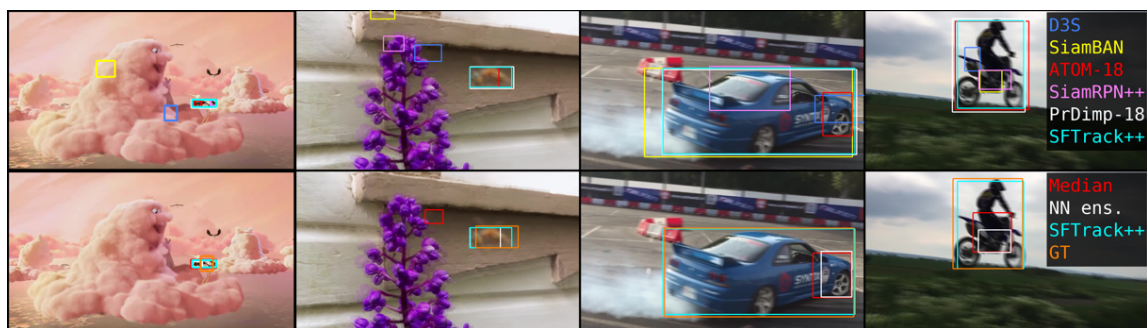


Figure 2: Qualitative results. We compare in the first line SFTrack++ with the input from individual methods. In the second line, we show the ground truth (in orange) and ensemble methods results that receive the same input as SFTrack++. We notice that even though the other ensembles fail to find a good bounding box, SFTrack++ manages to combine the input methods better, even in cases with a high variance among input methods.

In conclusion, SFTrack++ pre-registered proposal hypothesis validates through the proposed experimental protocol, with clear positive results.

7. Documented Modifications

There were several aspects where we need to deviate from the original protocol. Those aspects did not affect the core proposal and were mainly motivated by making the experiments less expensive in terms of computational cost, as detailed below.

Number of Benchmarks and Input Trackers. For each input tracker method considered, we run it in advance on all benchmarks (on training, valid, and test splits) to generate pre-processed input. We also generate the ground-truth bounding box segmentations for all benchmarks. This speeds up our training, making the overall training and testing self-contained, independent w.r.t. the input methods' code. We drop out the VOT2018 benchmark because its evaluation protocol consists of running a tracker on sub-videos, therefore we should have run all the input trackers code online, and this would have been too time-consuming. We resized each frame to keep its aspect ratio, having its maximum dimension of 480 pixels. For training, for each video in the tracking benchmarks, we used only a sample with 5 frames. The pre-processed data for one single tracker, on segmentation and tracking benchmarks, for training, valid, and testing splits takes ≈ 1 TB (without LaSOT). Since LaSOT has a very large number of frames, we decided to drop it out to make the experiment time manageable. We considered that if we test our proposal on 5 trackers and 5 benchmarks, our core concept of the proposal would not be affected and the results would be sufficiently general and conclusive. We chose the 5 trackers (out of the 10 in the proposal) that were the easiest to integrate with the PyTracking [pyt](#) (2019) framework.

NN_{feat2segm} **module.** Since the considered trackers were very different, we couldn't find a proper way to extract similar features from each tracker model and we decided to drop this ablation.

Bounding box regression. For extracting the bounding box coordinates out of the segmentation mask we use in all our experiments the region proposal from scikit-learn [Pedregosa et al. \(2011\)](#), with a 0.75 threshold for binarization. We did not perform an ablation study on bounding box regression because this would not be our contribution and did not influence our core proposal, this solution for bounding box regression being good enough to emphasize what we followed in our approach.

References

- PyTracking, 2019. URL "<https://github.com/visionml/pytracking>".
- Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *ISBI*, 2019.
- Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip H. S. Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, 2016a.
- Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In Gang Hua and Hervé Jégou, editors, *ECCV Workshops*, 2016b.
- Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *ECCV*, 2020.
- Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.
- Elena Burceanu and Marius Leordeanu. Learning a robust society of tracking parts using co-occurrence constraints. In *ECCV Workshops*, 2018.
- Elena Burceanu and Marius Leordeanu. A 3d convolutional approach to spectral object segmentation in space and time. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 495–501. ijcai.org, 2020. doi: 10.24963/ijcai.2020/69. URL <https://doi.org/10.24963/ijcai.2020/69>.
- S. Caelles, K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. *CVPR*, 2017.
- Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *CVPR*, 2020.
- Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *CVPR*, 2020.
- Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: efficient convolution operators for tracking. In *CVPR*, 2017.

- Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: accurate tracking by overlap maximization. In *CVPR*, 2019.
- Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020.
- Chris H. Q. Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *ICDM*, 2001.
- Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019.
- Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017.
- Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. *CVPR*, 2020a.
- Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. *CoRR*, 2020b.
- Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman P. Pflugfelder, Luka Cehovin Zajc, Tomas Vojır, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, Gustavo Fernandez, and et al. The sixth visual object tracking VOT2018 challenge results. In *ECCV 2018 Workshops*, 2018.
- Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. *ICCV*, 2005.
- Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018.
- Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- Alan Lukezic, Jiri Matas, and Matej Kristan. D3S - A discriminative single shot segmentation tracker. In *CVPR*, 2020.
- Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. In *AISTATS*, 2001.
- Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for UAV tracking. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016.

- Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018.
- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.
- Eunbyung Park and Alexander C. Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *ECCV*, 2018.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.
- F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.
- Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- Sudeep Sarkar and Padmanabhan Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. *PAMI*, 2000.
- Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. In *PAMI*, 2000.
- Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *MICCAI*, 2017.
- Ran Tao, Efstratios Gavves, and Arnold W. M. Smeulders. Siamese instance search for tracking. In *CVPR*, 2016.
- Andrea Torsello, Samuel Rota Bulò, and Marcello Pelillo. Grouping with asymmetric affinities: A game-theoretic perspective. In *CVPR*, 2006.
- Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandrar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTs: multi-object tracking and segmentation. In *CVPR*, 2019.
- Paul Voigtlaender, Jonathon Luiten, Philip H. S. Torr, and Bastian Leibe. Siam R-CNN: visual tracking by re-detection. In *CVPR*, 2020.
- Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *CVPR*, 2020.

- Pei Wang and Albert C. S. Chung. Focal dice loss and image dilation for brain tumor segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer F. Syeda-Mahmood, and et al., editors, *MICCAI*, 2018.
- Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.
- Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.
- Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, 2020.
- Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.
- Stella X. Yu and Jianbo Shi. Grouping with directed relationships. In *EMMCVPR 2001*, 2001.
- Zhipeng Zhang and Houwen Peng. Ocean: Object-aware anchor-free tracking. In *ECCV*, 2020.