Fast Video Object Segmentation with Temporal Aggregation Network and Dynamic Template Matching

Xuhua Huang[1*] Jiarui Xu[1*] Yu-Wing Tai[2] Chi-Keung Tang[1]
[1]The Hong Kong University of Science and Technology [2]Tencent
xhuangat@ust.hk jxuat@ust.hk yuwingtai@tencent.com cktang@cs.ust.hk

Abstract

Significant progress has been made in Video Object Segmentation (VOS), the video object tracking task in its finest level. While the VOS task can be naturally decoupled into image semantic segmentation and video object tracking, significantly much more research effort has been made in segmentation than tracking. In this paper, we introduce "tracking-by-detection" into VOS which can coherently integrates segmentation into tracking, by proposing a new temporal aggregation network and a novel dynamic time-evolving template matching mechanism to achieve significantly improved performance. Notably, our method is entirely online and thus suitable for one-shot learning, and our end-to-end trainable model allows multiple object segmentation in one forward pass. We achieve new state-of-the-art performance on the DAVIS benchmark without complicated bells and whistles in both speed and accuracy, with a speed of 0.14 *second per frame and* $\mathcal{J}\&\mathcal{F}$ *measure of* 75.9% *respectively. Project page is available at* https://xuhuaking.github.io/Fast-VOS-DTTM-TAN/.

1. Introduction

Video object segmentation (VOS) is a fine-grained labeling problem aiming to find pixel-level correspondence across frames in a given video, which has broad range of applications including surveillance, video editing, robotics and autonomous driving. In this work, we focus on the semi-supervised VOS setting in which the ground-truth segmentation masks of the target objects in the first frame are given. The task is then to automatically predict the segmentation masks of the target objects for the rest of the video. With the recent advances in deep learning and the introduction of the DAVIS datasets [44, 45], tremendous progress has been made in tackling this semi-supervised VOS task. Notwithstanding, existing state-of-the-art methods are heavily biased toward semantic segmentation in their design and thus they do not leverage the advantages of excellent tracking solutions. We believe that the VOS task can be naturally decoupled into image semantic segmentation and video object tracking:

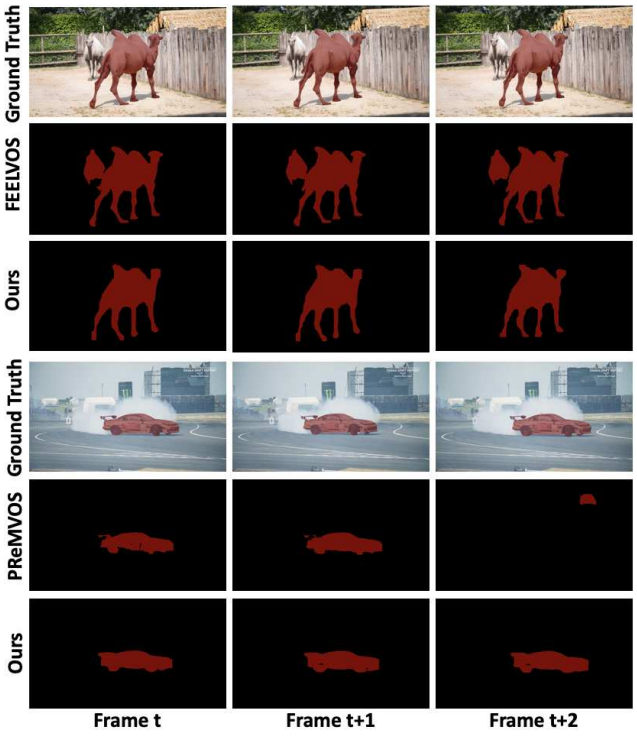

Figure 1. **Qualitative comparisons** on the DAVIS validation set. The camel example shows the case when FEELVOS [48] misjudges pixels from another object. The racing car example shows that PReMVOS [38] switches to another object between two consecutive frames. Both FEELVOS and PReMVOS are respectively the top 2 entries on the DAVIS benchmark.

In semantic segmentation, most works [8, 9, 10, 11] are mainly based on Fully Convolutional Networks [37]. Many VOS pipelines [29, 4, 49, 1, 48, 52, 41] exploit these architectures to produce a segmentation map. However, semantic segmentation is not instance sensitive. In complex scenarios where many different instances share the same semantics (e.g., pedestrians, vehicles and cases in Figure 1), these methods based on semantic segmentation may fail to track individual objects consistently. This has prompted us

*Equal contribution. This research is supported in part by Tencent and the Research Grant Council of the Hong Kong SAR under grant no. 1620818.

to revisit the whole pipeline from the perspective of Multiple Object Tracking (MOT).

In MOT, most recent approaches [54, 30, 47, 14, 47, 14, 3, 62, 56, 2] have adopted the popular tracking-by-detection model, where objects are first localized in each frame and then associated across frames. This model benefits from the rapid progress in the field of object detection [19, 46, 24, 33, 35, 16, 6], and has led to several popular benchmarks over the past few years, i.e., MOT15∼17 [31, 39]. Such a decoupled pipeline also makes it easy to extend or upgrade the tracking system with latest detection techniques.

Our work is motivated by the recently proposed tracker [2] which consists of further extension of the "tracking-by-detection" model without exploiting tracking annotation. Under the semi-supervised setting, even though the first frame ground-truth is given, tracking annotation is still unavailable. We fine-tune the detector with the first frame ground-truth segmentation mask for domain adaptation, whose iterations are proportional to the length of the sequence, and then perform tracking through the rest of the video. We further extend the tracker with our novel Temporal Aggregation Network and Dynamic Time-evolving Template Matching mechanism, both of which make the tracking pipeline more robust to occlusion and appearance changes.

Many leading methods rely on various overly-engineered design and/or heavy modules to improve their system performance while sacrificing run times. For example, PReMVOS [38], the winner of 2018 DAVIS challenge and current champion on DAVIS benchmark, adopts a total of four neural networks, optical flow and merging mechanism, and as a result it takes around 38 seconds to segment one single frame. DyeNet [32], another leading method in VOS, integrates FlowNet [28] for utilizing optical flow information and Bi-directional Recurrent Neural Network (RNN) for mask propagation, leading to a speed of around 2.4 seconds per frame. Though some of these methods can generate excellent results, the long run times and/or high complexity of their system prevent them from practical deployment or extension with advancement of relevant techniques.

In this paper, we present a simple, fast and high performance new baseline for VOS. Instead of formulating video object segmentation as mainly a segmentation problem, we propose to tackle the problem as a tracking problem with segmentation in fine-grained level. The proposed novel Temporal Aggregation Network and Dynamic Time-evolving Template Matching can be easily integrated into this simple pipeline to boost performance.

Our **contributions** can be summarized as follows:

- We present a simple, easy-to-extend and end-to-end trainable pipeline for VOS task by introducing the "tracking-by-detection" model into VOS which supports multiple object segmentation in one forward pass. To the best of our knowledge, we are the first to introduce "tracking-by-detection" model into video object segmentation as an online method with strong results and high speed;
- We contribute a novel Temporal Aggregation Network and Dynamic Time-evolving Template Matching mechanism, which are entirely online and naturally fit into one-shot learning to boost performance;
- We achieve a new state-of-the-art performance in both speed and accuracy without complicated bells and whistles on the DAVIS benchmark, with a speed of 0.14 second per frame and $\mathcal{J}\&\mathcal{F}$ mean score of 75.9%.

2. Related Work

Semantic Segmentation Recently, state-of-the-art semantic segmentation frameworks based on the fully convolutional network (FCN) [37] have made remarkable progress. Deeplab [8, 9, 10, 11] and context based methods [59, 50, 20, 60, 27] have further improved the performance over the years. However, benchmarks in semantic segmentation [15, 61] are used to evaluate full image semantic labeling task, while VOS belongs to instance segmentation of one or multiple specific target objects. For semantic segmentation, pixels in the same semantics are labeled with same category with no instance information. As for instance segmentation, the combination of both instance detection and pixel-level segmentation are required. The misalignment of these two tasks indicates that the design and method in semantic segmentation may not directly applicable in video object segmentation. Notwithstanding, with this intrinsic difference, many leading entries [29, 4, 49, 1, 48, 52, 41] in DAVIS benchmark still utilize this pipeline and adopt the architecture of semantic segmentation, which as a result make different instances vulnerable to switch and/or drift as shown in Figure 1.

Tracking-by-Detection Model Recent multiple object tracking (MOT) methods are mostly based on the tracking-by-detection model. As object detection has been well studied on its own, the main focus of MOT is on the data association problem. In MOT Benchmark [31, 39], public detectors are shared among all leading entries [47, 56]. This makes all pipelines designs adopting this model easy to extend with the latest detection approaches [19, 46, 24, 33, 35, 16, 6]. Even though MOT and VOS Benchmark settings differ in many aspects, the same intrinsic key is robust tracking ability. However, surprisingly only a few approaches formulate the VOS problem using this model and they do not come with online setting. While PReMVOS [38] may be regarded as some kind of close proximate to the tracking-by-detection model, its design is very complex and heavy, involving optical flow, multiple networks, and the framework cannot be trained end-to-end. Our method adopts the tracking-by-detection model which explicitly introduces detection and association into video object segmentation.

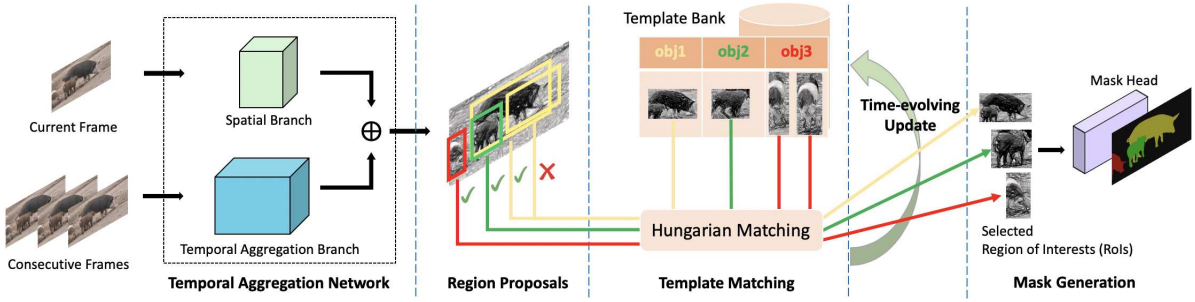

Figure 2. Overview of our pipeline.

Compared with PReMVOS, our approach is suitable for online setting, easy to extend and end-to-end trainable.

Video Object Segmentation Video Object Segmentation (VOS) aims for joint segmentation and tracking. The recently released DAVIS benchmarks [44, 45, 5] have contributed significantly in pushing the frontier of the relevant state-of-the-arts. However, as noted in [48], many of the methods do not fulfill the design goals of being robust, simple, fast and end-to-end trainable. For instance, as the winning entry of DAVIS 2018 Challenge [5], PReMVOS [38] exploits 4 different neural networks working in tandem, optical flow wrapping on image space and complicated merging algorithm, taking more than 38s for processing a given frame. The second runner up CINM [1] uses Markov Random Field (MRF) and iterative inference algorithm. DyeNet [32] incorporates template matching into a re-identification network; however optical flow warping and RNN are used in the mask network, thus making it training complicated and computationally demanding. Recently proposed FEELVOS [48] is an extension of MaskTrack [43] which uses previous frames predictions as input for the current frame, which is prone to accumulation error during tracking. Moreover, due to the absence of shared design model, many methods have limited practical usage and are not easy to extend. Our introduction and extension of "tracking-by-detection" model aims to establish a new, simple and strong baseline for video object segmentation.

3. Method

Refer to Figure 2. The feature map of the current frame is extracted through our Temporal Aggregation Network (TAN) which coherently combines spatial and temporal information. TAN aggregates neighboring frame spatial features across time to enrich the current frame feature. On top of the aggregated feature map from TAN, a Region Proposal Network (RPN) is applied to extract region proposals which are likely to contain target objects. With region proposals in the current frame, Hungarian matching [40] is performed between template features in the bank and current detected object features. Initially, the template features are just the features of the first frame ground truth objects. After matching, the most confident region proposal for each object will be chosen and passed to the Mask Head for final segmentation generation. At the same time, if large appearance change occurs within the chosen proposals, the new proposals would serve as new templates for future frames.

We first introduce how we incorporate tracking-by-detection model into the video object segmentation task. Then we will present our proposed Temporal Aggregation Network and Dynamic Time-evolving Template Matching.

3.1. Tracking by Detection

The goal of video object segmentation (VOS) is similar to multi-object tracking (MOT), which is to predict the trajectories of multiple objects at a fine-grained level over time. We denote the set of trajectories $\mathbf{T} = \{\mathbf{T}_i\}_{i=1}^{N}$, where the trajectory of the i^{th} object can be represented by a series of bounding boxes together with the masks within, denoted by $\mathbf{T}_i = \{\mathbf{b}_i^t, \mathbf{m}_i^t\}_{t=1}^{T}$, $\mathbf{b}_i^t = [x_i^t, y_i^t, w_i^t, h_i^t]$; x_i^t and y_i^t denote the center location of the target i at frame t; w_i^t and h_i^t denote respectively the width and height of the target object i; $\mathbf{m}_i^t$ denotes the corresponding foreground segmentation mask within corresponding bounding box. In MOT usually only $\{\mathbf{b}_i^t\}_{t=1}^{T}$ is needed, while VOS is more fine-grained and should produce $\{\mathbf{m}_i^t\}_{t=1}^{T}$. Bounding box representation is popular in recent detection pipelines [46, 24] while some non-box based methods [12] have also been proposed. We mainly adopt box based detection pipelines in this paper but note that our method is also compatible with non-box based pipelines.

To output segmentation mask, a light-weight Fully Convolutional Network is attached to the detector; foreground/background segmentation is performed on each detected bounding boxes. For simplicity, we ignore the $\{\mathbf{m}_i^t\}_{t=1}^{T}$ which can be easily generated from $\{\mathbf{b}_i^t\}_{t=1}^{T}$. Under semi-supervised setting of DAVIS Benchmark [44, 45, 5], the tracking targets are given in the first frame. We denote tracking targets as $\{\mathbf{t}_i\}_{i=1}^{N}$.

Our method follows the online tracking-by-detection model [54], which first detects multiple objects in each frame and then associates their identities across frames. Following [4, 29] we first train the whole detector offline with the training set in order to adapt the detector to the DAVIS dataset domain. Before inference, one-shot learning is performed with the given ground-truth segmentation for a few iterations to reduce false alarm. Note that one advantage of our training strategy is that no tracking-specific training is required, which implies that our method can be

readily extended to various applications.

During inference, different from [2] and for simplicity we do not exploit bounding box regressor for tracking. Instead, simple Intersection over Union (IoU) metric is adopted. For segmentation based approaches, IoU is hard to be incorporated into the pipeline since there is no explicit bounding box or instance. Our method first detects multiple objects in each frame and then associates their identities across frames. The detected boxes at frame t with confidence score greater than σ_{det} are denoted as $\{\mathbf{b}_j^t\}_{j=1}^{N_t} = D_t$. Note that N_t is not necessarily equal to N due to false positives and false negatives output of the detector. At $t = 0$, our tracker initializes the tracklet from the first frame bounding boxes $\{\mathbf{t}_i\}_{i=1}^{N} = \{\mathbf{b}_j^0\}_{j=1}^{N_0} = D_0$. At frame t, a bipartite graph is constructed based on location similarity IoU, the weight ω_{ij}^{loc} between ith target and jth detected object in the current frame is defined as the IoU of $\mathbf{t}_i$ and $\mathbf{b}_j^t$. After Hungarian matching [40] is performed, the target $\mathbf{t_i}$ will be updated with $\mathbf{b}_k^t$ and added to $\mathbf{T}_i$ if matched. When running the assignment process in a frame-by-frame manner the object trajectories are produced.

This naive tracking-by-detection pipeline under the setting of semi-supervised video object segmentation will be extended in following sections.

3.2. Temporal Aggregation Network

Figure 3 illustrates the Temporal Aggregation Network (TAN) with details to be described in the following.

Many methods in object tracking have exploited temporal information to facilitate tracking. However, most of them require optical flows to obtain pixel-wise alignment [64, 63] which can be quite computationally extensive. Inspired by the recent progress in large-scale action recognition benchmarks [7, 22], we propose a novel Temporal Aggregation Network to incorporate backbones in image classification and video recognition, which align well with the tracking-by-detection model as well as detection in image and tracking in video.

Image Backbone Most object detection methods have exploited ImageNet [17] pretrained backbone for faster convergence as studied in [23]. We use ResNet [25] as the standard setting similarly done in other detection approaches. We denote the outputs of stage 3, 4, 5 as c3, c4, c5 respectively. The input of Image Backbone is the key frame where objects are to be detected.

Video Backbone Inspired by recent progress in Human Action Recognition [7, 51, 18], we believe 3D convolution is effective in utilizing temporal information across consecutive frames, although no previous works have attempted to utilize this technique to tackle VOS task. Note that while optical flow approaches in [38, 32] utilize explicit correspondence, 3D convolution network can directly learn temporal patterns from an RGB stream. According to [7], I3D performs best against other methods such as CNN+LSTM counterpart. Therefore, we adopt I3D in our video backbone and at the same time, similar to [51], $\text{I3D}_{3\times1\times1}$ and $\text{I3D}_{1\times3\times3}$ are used in order to reduce the computation cost of 3D convolution. We denote the outputs of stage 3, 4, 5 as i3, i4, i5 respectively. The previous α frames and current key frame are concatenated along the temporal axis resulting in a 3D tensor of size $(\alpha + 1) \times H \times W$ which is then fed into the video backbone.

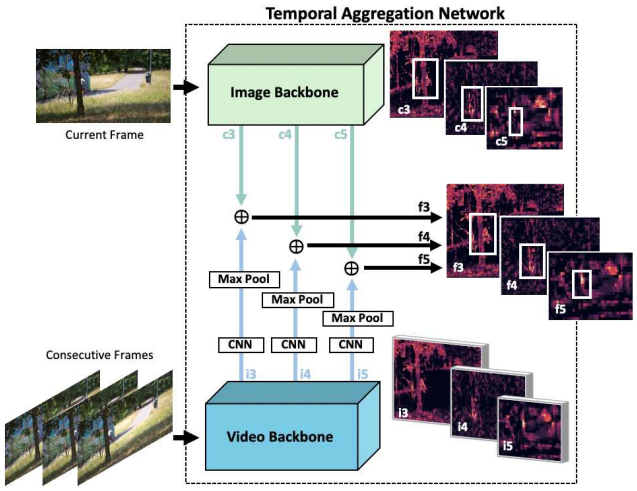

Figure 3. Illustration of Temporal Aggregation Network. Though in c3, c4, c5 from Image Backbone the activation around the target object is weak, after aggregation by Video Backbone, the final output f3, f4, f5 can receive proper activation around the target.

Temporal Aggregation Fusion Inspired by Slow-Fast [18], the separation and fusion of fast/slow video stream is a good practice to make use of temporal information while preserving key frame feature. As conv-add style in [18], the new feature map f_i is computed by $\text{c}_i + \mathcal{N}(\text{i}_i)$ for stages 3, 4, 5, where $\mathcal{N}$ denotes a small CNN to fuse features with Max Pooling to compress information along the temporal dimension so that the two branches have the same size at each stage. f_i will be used in the subsequent detection networks. Unlike RNN based methods [32, 47] which require sophisticated training schedule, our network is feed-forward and can be jointly trained. As shown in Figure 5 in the experimental section, by incorporating temporal information, our system can better handle occlusion.

3.3. Dynamic Time-evolving Template Matching

Obviously the naive IoU based tracking algorithm is not robust to large camera movement or object deformation. As pointed out in [2, 38], re-identification is also essential to object tracking. Thus we propose the novel Dynamic Time-evolving Template Matching (DTTM) to tackle the re-identification problem. Figure 4 summarizes the method.

Instead of cropping bounding box regions from RGB image space, we prefer the feature produced by the backbone since it is encoded with high-level semantic meaning. Let $\mathcal{A}(\mathbf{b}^t)$ denote the appearance feature in bounding box $\mathbf{b}$ extracted from the backbone feature map of frame t which is a

high dimensional vector. The appearance similarity weight term is defined as

$$\omega_{ij}^{\text{app}} = \frac{\mathcal{A}(\mathbf{t}_i) \cdot \mathcal{A}(\mathbf{b}_j^t)}{\|\mathcal{A}(\mathbf{t}_i)\| \cdot \|\mathcal{A}(\mathbf{b}_j^t)\|} \tag{1}$$

As pointed out in [47, 56], location cue and appearance cue are essential features in multiple object tracking. Therefore the bipartite graph is constructed using $\omega_{ij} = \omega_{ij}^{\text{loc}} + \omega_{ij}^{\text{app}}$ in order to take both cues into account, where ω_{ij}^{loc} has been defined in Section 3.1.

In ideal scenarios, the learned appearance feature from the first frame is sufficient for tracking the rest. However, a constant template clearly does not work in practice especially in long-term tracking, since the target objects may undertake many appearance changes due to deformation, occlusion, etc. So a proper design for **time-evolving** template is essential, which leads to our online updating template appearance features during the tracking progress.

Moving average is one of the widely adopted approaches for template update across temporal axis. The single hyper parameter $momentum$ controls the proportion of features to be updated or preserved, which is extremely sensitive to the frame rates of different videos. The static averaging process is also obviously sub-optimal due to accumulation error and feature blurring. We propose to update the template feature more **dynamically** in a discrete manner.

DTTM Full algorithm detailing our matching strategy in pseudocodes can be found in the supplementary material. Initially, the template bank for each target i is constructed as $bank_i = \{\mathbf{t}_i\}$. The matching result is obtained by performing linear assignment between $\mathbf{t}_i \in \cup_{i=1}^{N} bank_i$ and $\mathbf{b}_j^t \in \{\mathbf{b}_j^t\}_{j=1}^{N_t}$ at each frame t by computing $\omega_{ij}^{\text{loc}} + \omega_{ij}^{\text{app}}$. Given foreground confidence score $\text{conf}(\mathbf{b}_j^t)$ greater than threshold σ_{conf}, when $\mathbf{t}_i$ and $\mathbf{b}_j^t$ is a match but their appearances are not similar $\omega_{ij}^{\text{app}} < \sigma_{\text{app}}$, which indicates a drastic appearance change and thus a very likely situation of losing track in the upcoming frames, our DTTM will initialize a new template from the latest matched detection $\mathbf{b}_j^t$ for the target object $\mathbf{t}_k$ and add it into corresponding feature bank, which results in an extended template bank for the future assignment.

To avoid overflow, the least frequently used template will be removed from the $bank_i$ when the number of templates is larger than some threshold. Both the initial target feature and updated feature can be considered as templates during later matching process. Our system can thus detect potential accumulation error early and thus prevent its adverse effect by updating with high confidence new template feature. Figure 4 illustrates how DTTM can effectively address the deformation problem in VOS, where not only the latest but also previous templates are considered so that abrupt deformation can be well handled. Note that this matching mechanism is entirely online and time-evolving unlike [38], namely, in our DTTM no future information is required and the template bank keeps evolving over time.

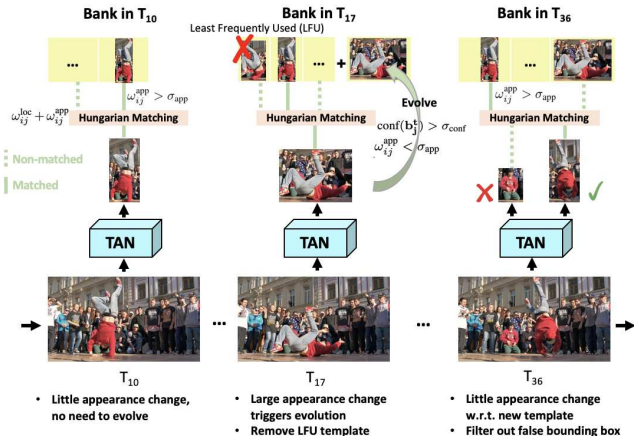

Figure 4. This figure demonstrates an effective case of our DTTM on handling deformation problem. For visualization here we use image to replace feature vector after TAN stage.

4. Experiments

4.1. Dataset and Evaluation Metrics

We use DAVIS benchmarks [44, 45, 5] to evaluate our method. These benchmarks are challenging, with large variety in scenes, multiple heterogeneous objects and occlusion, and are widely used in evaluating video object segmentation methods.

DAVIS 2016 DAVIS 2016 comprises of a total of 50 sequences, 3455 annotated frames, all of which were captured at 24fps and Full HD 1080p spatial resolution. Since computational complexity is a major bottleneck in video processing, each sequence has a short temporal extent (about 2–4 seconds) while including all major challenges typically found in longer video sequences

DAVIS 2017 After releasing DAVIS 2016 [44], several strong methods have been proposed making the top performance on DAVIS 2016 saturated. DAVIS 2017 was released which is an extension of the former version and a more challenging benchmark. Overall, the new dataset consists of 150 sequences, totaling 10459 annotated frames and 376 objects. One major difference from DAVIS 2016 is that in DAVIS 2017, multiple object tracking is introduced in video object segmentation. In DAVIS 2016, only single objects are annotated for each frame while in DAVIS 2017, multiple object masks in the same frame are annotated, which makes the task more challenging as complex interaction among multiple target objects can cause occlusion and thus change of topology and appearance. Moreover, as Figure 1 shows, many target objects are in the same category and have very similar appearance. Another important challenge in DAVIS 2017 is that target object category in one video can become background in another video, thus demanding high discriminative ability of the tracker to adapt to different target objects given only the first frame annotation. Note that DAVIS 2017 is also the dataset for DAVIS 2019 Challenge, which we will use in the following unless otherwise specified.

4.2. Implementation Details

We use Faster R-CNN [46] as the detector and ResNet-50 [25] as our default backbone unless otherwise specified. The spatial branch backbone is initialized with weights given by ImageNet [17] classification and the detector is pretrained on COCO [36] as done in [38]. As for the temporal aggregation branch, we use 8-frame input I3D baseline provided by Nonlocal Network [51]. In order to incorporate multiple scales during tracking, feature pyramid network (FPN) [34] is adopted in the backbone to merge high-level semantic information from deeper coarse feature map with shallow fine-grained feature map. Following [34], the RPN anchors span 5 scales (feature is extracted from different levels of FPN outputs) and 3 aspect ratios [0.5, 1.0, 2.0]. Since category classification is unavailable in the DAVIS dataset, the last linear classification layer of R-CNN is replaced by a simple fully connected layer for foreground/background classification. This class agnostic R-CNN is more suitable for semi-supervised setting and more general for real applications. Vanilla Fully Convolution Network [37] is adopted as the segmentation head as default. As in Fast R-CNN [21], an RoI is considered positive if its IoU with the ground-truth box is at least 0.5 and negative otherwise. The segmentation is only performed on positive RoIs during both training and testing.

Training To overcome the domain gap between the pretrained dataset and DAVIS, we first train the entire detection network on DAVIS training set. The input images are resized such that their shorter side is 800p which is similarly done in [46] unless otherwise specified. We train on 8 GPUs with 2 images per GPU (effective mini batch size of 16). All layers of backbone except for c1 and c2 of backbone are jointly fine-tuned with detection annotation. Due to the limited batch size, all batch normalization layers are frozen so that the statistics of mean and variance are unchanged during training. Unlike stage-wise training with respect to RPN in [24], end-to-end training as in [34] is adopted in our implementation which yields better results. All models are trained for 100k iterations using synchronized SGD with a weight decay of 0.0001 and momentum of 0.9. The learning rate is initialized to 0.002, which decays by a factor of 10 after 70k iterations and 90k iterations. Other choices of hyper-parameters also follow the setting in [46].

Inference At test time, we mostly follow the setting of Faster-RCNN in [34]. The box prediction is done on RPN proposals, followed by non-maximal suppression [19]. The segmentation head is then applied to the detected bounding boxes above threshold σ_{conf} only. The instance segmentation results are merged into a single segmentation map by ranking corresponding bounding box confidence scores.

4.3. Ablation Study

The ablation study is performed on the DAVIS 2017 dataset. The outputs are resized to 480p for evaluation.

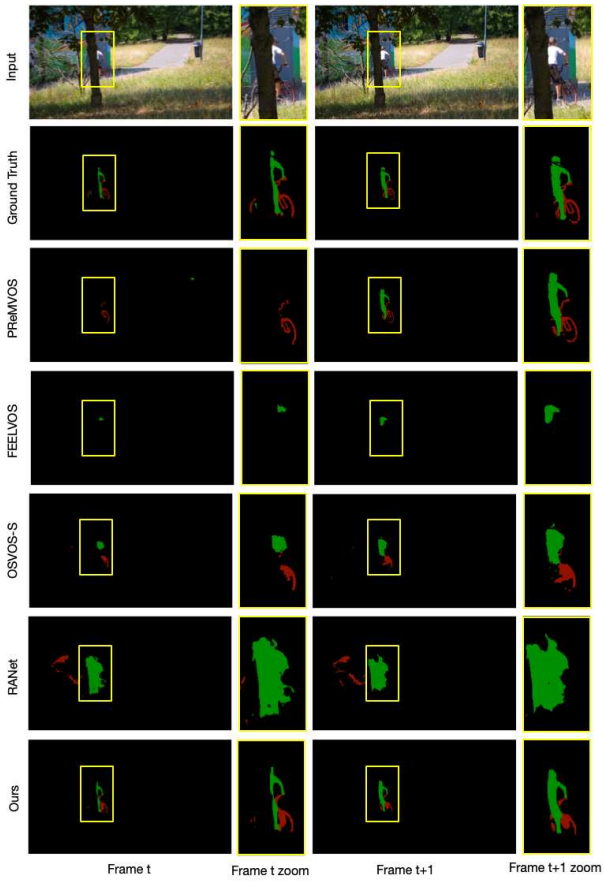

Figure 5. **Disocclusion of man riding bike**. Results of Temporal Aggregation Network compared with Ground Truth, PReMVOS [38], FEELVOS [48], OSVOS-S [4] and RANet [52].

res	TAN	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
800		45.3	66.5	51.1	35.4	59.4	36.9
800	✓	$\mathbf{46.3}_{\uparrow 1.0}$	$\mathbf{70.9}_{\uparrow 4.4}$	$\mathbf{52.3}_{\uparrow 1.2}$	$\mathbf{36.7}_{\uparrow 1.3}$	$\mathbf{60.9}_{\uparrow 1.5}$	$\mathbf{38.1}_{\uparrow 1.2}$
600		45.0	66.5	51.0	35.3	59.0	37.2
600	✓	$\mathbf{46.3}_{\uparrow 1.3}$	$\mathbf{71.0}_{\uparrow 4.5}$	$\mathbf{52.5}_{\uparrow 1.5}$	$\mathbf{36.2}_{\uparrow 0.9}$	$\mathbf{61.0}_{\uparrow 2.0}$	$\mathbf{38.0}_{\uparrow 0.8}$
400		44.5	66.6	50.9	35.2	58.9	**37.5**
400	✓	$\mathbf{46.6}_{\uparrow 2.1}$	$\mathbf{71.2}_{\uparrow 4.6}$	$\mathbf{52.9}_{\uparrow 2.0}$	$\mathbf{35.8}_{\uparrow 0.6}$	$\mathbf{62.9}_{\uparrow 4.0}$	$36.9_{\downarrow 0.6}$

Table 1. Results of Temporal Aggregation Network (TAN) on DAVIS validation set. This table presents results under different input resolutions when comparing with the baseline (i.e. no TAN).

4.3.1 Temporal Aggregation Network

We report the standard object detection and instance segmentation COCO metrics including AP, AP50, AP75 for both bounding boxes and segmentation masks. Figure 5 shows that our Temporal Aggregation Network is more robust to occlusion compared with other methods.

Different Input Resolutions Although high-resolution input evidently benefits the localization ability of a network, it is not always applicable due to hardware constraints. Table 1 shows that even though the computation cost increases quadratically when input resolution increases, the performance gain is only marginal and saturates very soon (e.g. first column, $AP^{bbox/mask}$ 45.3/35.4 at 800 res. versus

method	optical flow	backbone	input	AP^{bbox}_{50}
DFF [64]	✓	ResNet-101	Continuous	73.0
FGFA [63]	✓	ResNet-101	Continuous	76.8
Ours		ResNet-50	Duplicate	76.0
Ours		ResNet-50	Continuous	$\mathbf{78.2}_{\uparrow \mathbf{2.2}}$

Table 2. Results of Temporal Aggregation Network (TAN) on ImageNet VID validation set. Duplicate denotes simply stacking the same frame to fit the network input size, while continuous means both current and neighbor frames are fed into TAN.

$AP^{bbox/mask}$ 44.5/35.2 at 400 res.). By aggregating temporal information, even low-resolution input can achieve better performance (e.g. first column, $AP^{bbox/mask}$ 45.3/35.4 at 800 res. versus $AP^{bbox/mask}$ 46.6/35.8 at 400 res. with TAN). It is also worth noting that the baseline method performance drops significantly under low-resolution setting, while from different input resolutions TAN achieves consistently higher performance, which indicates that effectively aggregating temporal information across frames can complement loss of resolution due to compression to some extent.

More Challenging Dataset Since the number of target object in DAVIS dataset is quite limited (1 in DAVIS 2016, 1 to 3 in DAVIS 2017), we also demonstrate the effectiveness of our temporal aggregation network on ImageNet VID dataset [17] which is larger (3962 training sequences, 555 validation sequences) and more complicated (1 to 10 target objects from 30 categories). Table 2 presents our quantitative results. Our method's performance is on par with popular optical flow feature alignment methods with stronger backbones. In addition, using continuous frames as input can boost up performance by 2 points compared with duplicate frames, which again demonstrates that temporal information is effectively utilized by our proposed Temporal Aggregation Network.

4.3.2 Dynamic Time-evolving Template Matching

Besides no matching and IoU matching, we also compare our dynamic template matching with naive moving average with momentum mnt, in which matching is based on $\mathcal{A}(\mathbf{t}_i) = (1 - mnt)\mathcal{A}(\mathbf{t}_i) + mnt\mathcal{A}(\mathbf{b}_j)$ where $\mathbf{t}_i$ and $\mathbf{b}_j$ is a match. As Table 3 shows, by setting $\theta_{\text{conf}} = 0.5$ and sliding θ_{app}, our dynamic template matching outperforms the baseline by a large margin. Note that for the DTTM module, if the threshold becomes too low (e.g. 0.3), it may introduce misleading templates. On the other hand, if the threshold becomes too high (e.g. 0.7), it may filter out useful templates. Either case will lead to performance drop. Note that even though moving average can also improve the performance, it is worse than DTTM in general. We argue that due to the high frame rate of DAVIS dataset, it is hard for naive moving average to model appearance changes which makes it more prone to accumulation error.

Qualitative comparison with leading approaches are provided in Figure 6. Multiple major tracking challenges are presented in this example, such as topology change, occlusion, similar texture pattern and semantic. Embedding and

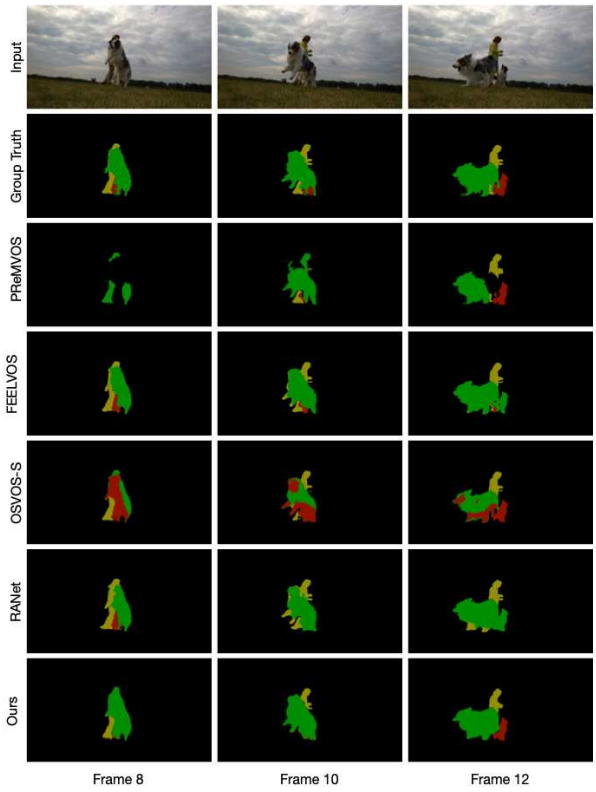

Figure 6. **Large deformation (foreground dog) with disocclusion of a similar but different instance (background dog)**. Results of Dynamic Time-evolving Template Matching compared with Ground Truth, PReMVOS [38], FEELVOS [48], OSVOS-S [4] and RANet [52].

method	setting	$\mathcal{J}\&\mathcal{F}$-Mean	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
IoU match		69.2	65.8	72.6
moving avg	$mnt = 0.2$	$70.1_{\uparrow 0.9}$	$66.8_{\uparrow 1.0}$	$73.4_{\uparrow 0.8}$
moving avg	$mnt = 0.3$	$\mathbf{70.8}_{\uparrow \mathbf{1.6}}$	$\mathbf{67.5}_{\uparrow \mathbf{1.7}}$	$\mathbf{74.1}_{\uparrow \mathbf{1.5}}$
moving avg	$mnt = 0.5$	$69.4_{\uparrow 0.2}$	$66.1_{\uparrow 0.3}$	$72.8_{\uparrow 0.2}$
DTTM	$\theta_{\text{app}} = 0.3$	$70.5_{\uparrow 1.3}$	$67.3_{\uparrow 1.5}$	$73.7_{\uparrow 1.1}$
DTTM	$\theta_{\text{app}} = 0.5$	$\mathbf{71.7}_{\uparrow \mathbf{2.5}}$	$\mathbf{68.5}_{\uparrow \mathbf{2.7}}$	$\mathbf{74.9}_{\uparrow \mathbf{2.3}}$
DTTM	$\theta_{\text{app}} = 0.7$	$70.6_{\uparrow 1.4}$	$67.4_{\uparrow 1.6}$	$73.9_{\uparrow 1.3}$

Table 3. Results of moving average and Dynamic Time-evolving Template Matching (DTTM) on DAVIS validation set.

segmentation based approaches like [48, 4, 52] failed to distinguish two dogs and even produce shattered segmentation. Moreover [38] failed to incorporate drastic topology (pose) change of the jumping dog resulting in false tracking, while DTTM can easily delineate targets in similar appearance and keep tracking of rapidly deforming objects.

4.3.3 Segmentation Head

Different off-the-shelf design choices of segmentation head have also been investigated. We deploy two types of skip connection, *cascade* and *join*, where *cascade* sequentially adds the output of each layer to the input of next layer as residue, and *join* simply aggregates all layer outputs by ad-

Figure 7. Qualitative results on DAVIS validation set. Several challenging cases are presented, such as occlusion, deformation, zoom in/out, to demonstrate robustness of our method.

head	skip connection	$\mathcal{J}\&\mathcal{F}$-Mean	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
FCN		67.9	64.2	71.5
FCN	✓*cascade*	$68.8_{\uparrow 0.9}$	$65.4_{\uparrow 1.2}$	$72.2_{\uparrow 0.7}$
FCN	✓*join*	$\mathbf{69.2}_{\uparrow \mathbf{1.3}}$	$\mathbf{65.8}_{\uparrow \mathbf{1.6}}$	$\mathbf{72.6}_{\uparrow \mathbf{1.1}}$

Table 4. Results of different types of segmentation heads on DAVIS validation set. These results are produced without using TAN and DTTM.

dition as final output. Table 4 shows the results of different segmentation heads. Note the simple skip connection has significant improvement over the vanilla FCN counterpart, which indicates that our method can be further benefited from advanced segmentation techniques.

4.4. Results on Benchmark

Table 5 tabulates the quantitative results in comparison with leading methods with qualitative results presented in Figure 7. We achieve state-of-the-art single model results in both speed and accuracy on DAVIS Benchmark with stronger backbone ResNeXt-101 [55] and deformable convolution [16], both contribute to further improving overall performance, which again justifies the advantage of our simple and easily extensible pipeline. Despite DAVIS dataset, many leading methods exploit YouTube-VOS [57] as pretraining. In Table 6 we demonstrated that our method achieved higher performance without computational demanding YouTube-VOS pretraining. The speed (t/s) reported in Table 6 considers that there are more than 2 objects on average on DAVIS-2017 dataset. STM [42] needs to handle them individually, while we could detect and track all the targets in one pass. We also reported a reasonable result on YouTube-VOS, 73.5% $\mathcal{J}\&\mathcal{F}$-Mean without carefully selecting hyperparameter.

5. Conclusion

Many leading VOS methods are overly complicated utilizing computationally-heavy modules or highly-engineered pipelines leading to limited practical usage. In this paper, we design a new and strong baseline that simultaneously achieves state-of-the-art speed and accuracy, by integrating the tracking-by-detection model into VOS,

Method	t/s	$\mathcal{J}\&\mathcal{F}$-Mean	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
PReMVOS [38]	37.6	**77.8/71.6**	**73.9/67.5**	**81.7/75.8**
OnAVOS [49]	26	63.6/52.8	61.0/49.9	66.1/55.7
FAVOS [13]	1.2	58.2/43.6	54.6/42.9	61.8/44.2
VideoMatch [26]	0.35	62.4/–	56.5/–	68.2/–
FEELVOS [48]	0.51	71.5/57.8	69.1/55.1	74.0/60.4
OSMN [58]	**0.28**	54.8/41.3	52.5/37.7	57.1/44.9
RGMP [53]	**0.28**	66.7/52.8	64.8/51.3	68.6/54.4
Ours	**0.14**	**75.9/65.4**	**72.3/61.3**	**79.4/70.3**

Table 5. Comparisons with state-of-the-art methods on the **validation/test-dev set** of DAVIS 2019 Challenge. t/s denotes running time per frame in seconds. The table demonstrates that our method achieves state-of-the-art performance in both speed and accuracy.

Method	t/s	YV	$\mathcal{J}\&\mathcal{F}$-Mean	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
FEELVOS [48]	0.51		69.1/54.4	65.9/51.2	72.3/57.5
		✓	71.5/57.8	69.1/55.2	74.0/60.5
STM [42]	0.32		71.6/-	69.2/-	74.0/-
		✓	81.7/72.2	79.2/69.3	84.3/75.2
Ours	0.14		75.9/65.4	72.3/61.3	79.4/70.3

Table 6. Comparisons with state-of-the-art methods on the **validation/test-dev set** of DAVIS-2017. YV denotes whether YouTube-VOS is used during training.

since VOS can be naturally decoupled into image semantic segmentation and video object tracking. With this design, our method is easy to extend because ongoing advancement in object tracking can further improve our method in future. With the introduction of multiple object segmentation in the DAVIS 2017 challenge, most leading methods at the time supporting multiple object segmentation required extra modification, whereas our method handles multiple object segmentation in one forward pass. On top of our design, we propose the novel Temporal Aggregation Network (TAN) and Dynamic Time-evolving Template Matching (DTTM), and their effectiveness have been demonstrated experimentally. Without bells and whistles, our method achieves a new state-of-the-art result on DAVIS benchmark for VOS. We hope our fast, practical and easy to extend pipeline will serve as a new baseline for future development targeting at higher efficiency, accuracy and scalability in VOS.

References

[1] Linchao Bao, Baoyuan Wu, and Wei Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2018. 1, 2, 3

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019. 2, 4

[3] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 2

[4] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 1, 2, 3, 6, 7

[5] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 3, 5

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 2

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018. 1, 2

[12] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019. 3

[13] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[14] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4836–4845, 2017. 2

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 2, 8

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 4, 6, 7

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019. 4

[19] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, Sept. 2010. 2, 6

[20] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 2

[21] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[22] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 4

[23] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019. 4

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 3, 6

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[26] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–70, 2018. 8

[27] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019. 2

[28] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[29] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. In *The DAVIS Challenge on Video Object Segmentation*, 2017. 1, 2, 3

[30] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2016. 2

[31] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942. 2

[32] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *European Conference on Computer Vision*, 2018. 2, 3, 4

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2, 6

[38] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[39] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 2

[40] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 3, 4

[41] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[42] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019. 8

[43] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2663–2672, 2017. 3

[44] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 1, 3, 5

[45] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 3, 5

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3, 6

[47] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311. IEEE, 2017. 2, 4, 5

[48] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 1, 2, 3, 6, 7, 8

[49] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation. In *The 2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops*, volume 5, 2017. 1, 2, 8

[50] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 4, 6

[52] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. *arXiv preprint arXiv:1908.06647*, 2019. 1, 2, 6, 7

[53] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2018. 8

[54] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*, pages 4705–4713, 2015. 2, 3

[55] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 8

[56] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019. 2, 5

[57] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 8

[58] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018. 8

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[60] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 2

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[62] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018. 2

[63] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–417, 2017. 4, 7

[64] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2017. 4, 7