TTVOS: Lightweight Video Object Segmentation with Adaptive Template Attention Module and Temporal Consistency Loss

ABSTRACT

Semi-supervised video object segmentation (semi-VOS) is widely used in many applications. This task is tracking class-agnostic objects by a given segmentation mask. For doing this, various approaches have been developed based on online-learning, memory networks, and optical flow. These methods show high accuracy but are hard to be utilized in real-world applications due to slow inference time and tremendous complexity. To resolve this problem, template matching methods are devised for fast processing speed, sacrificing lots of performance. We introduce a novel semi-VOS model based on a template matching method and a novel temporal consistency loss to reduce the performance gap from heavy models while expediting inference time a lot. Our template matching method consists of short-term and long-term matching. The short-term matching enhances target object localization, while long-term matching improves fine details and handles object shape-changing through the newly proposed adaptive template attention module. However, the long-term matching causes error-propagation due to the inflow of the past estimated results when updating the template. To mitigate this problem, we also propose a temporal consistency loss for better temporal coherence between neighboring frames by adopting the concept of a transition matrix. Our model obtains 79.5% J&F score at the speed of 73.8 FPS on the DAIS16 benchmark.

CCS CONCEPTS

• Computing methodologies → Video segmentation.

KEYWORDS

Semi-supervised video segmentation, video tracking, lightweight segmentation

1 INTRODUCTION

Video object segmentation (VOS) is essential in many applications such as autonomous driving, video editing, and surveillance system. In this paper, we focus on a semi-supervised video object segmentation (semi-VOS) task, which is to track a target in a pixel-wise resolution from a given annotated mask for the first frame.

For accurate tracking, many models have been developed, but it is hard to use the models in real-world environment due to tremendous computation. For example, one of popular method, online-learning, fine-tunes model parameters using the first frame image and the corresponding ground truth mask. This strategy makes the model more specialize in each video input, but, it requires additional time and memory for fine-tuning. Memory network method achieves high accuracy than any other approaches. They stacks multiple target memories and match the current frame with the entries. Therefore, the inference time and the required memories increase in proportion to the number of frames. To solve these problems, GC [16] conducted weighted-average to the multiple memories at each time frame for generating one global context memory. However, it needs an additional feature extraction step for updating the memory from the current estimated mask and the image. Also, we believe that it is not enough to directly comprehend spatial information since the size of global context memory much smaller than original spatial resolution size.

For increasing consistency of masks across frames, optical flow is one of the popular methods in low-level vision which has been applied in diverse video applications. In a video segmentation task, it propagates a given mask or features by computing pixel-wise trajectories or movements of objects. However, it is too demanding to compute exact flow vectors which contain excessive information for the segmentation task. For example, if we know the binary information of whether a pixel is changed into the foreground or background, we do not need an exact flow vector of each pixel.

The aforementioned methods have increased accuracy a lot, but they require heavy inference time and memory. The template matching approach resolves this problem by designing a target template from a given image and annotation. However, the accuracy is lower compared to other models because the matching method is too simple, and the template is hard to handle object shape variation.

In this paper, we propose an adaptive template matching method and a novel temporal consistency loss for semi-VOS. Our contributions can be summarized as follows: 1) We propose a new lightweight VOS model based on template matching method by combining short-term and long-term matching to achieve fast inference time and to reduce the accuracy gap from heavy and complex models. More specifically, in short-term matching, we compare the current frame’s feature with the information in the previous frame for localization.

Figure 1: The speed (FPS) vs accuracy (J&F score) on the DAVIS2016 validation set. Our proposed TTVOS achieves high accuracy with small complexity. HR/RN respectively denotes HRNet/ResNet50 for the backbone network.
2 RELATED WORK

Optical flow: Optical flow which estimates flow vectors of moving objects is widely used in many video applications [7, 12, 28, 31]. In the semi-VOS task, it aligns the given mask or features with the estimated flow vector. SegFlow [4] designed two branches, each for image segmentation and optical flow. The outputs of both branches are combined together to estimate the target masks. Similarly, FAVOS [17] and CRN [9] refined a rough segmentation mask by optical flow.

Online-learning: The online-learning method is training the model with new data in each inference iteration [14, 27, 45]. In the semi-VOS task, model parameters are fine-tuned in the inference stage with a given input image and a corresponding mask. Therefore, the model is specialized for the given condition of the clip [2, 20, 23]. However, fine-tuning causes additional latency in inference time. [26] resolved this issue by dividing the model into two sub-networks. One is a lightweight network that is fine-tuned in the inference stage for making a coarse score map. The other is a heavy segmentation network without the need for fine-tuning. This network enables fast optimization and relieves the burden of online-learning.

Memory network: The memory network constructs external memory representing various properties of the target. It was devised for handling long-term sequential tasks in the natural language processing (NLP) domain, such as the QA task [13, 30, 42]. STM [22] adopted this idea for the semi-VOS task by a new definition of key and value. The key encodes visual semantic clue for matching and the value stores detailed information for making the mask. However, it requires lots of resources because the amount of memory is increased over time. Furthermore, the size of memory is the square of the resolution of an input feature map. To lower this huge complexity, GC [16] does not stack memory at each time frame, but accumulate them into one, which is also of a smaller size than a unit memory of STM. They do not make a \( (hw \times hw) \) memory like [39, 47] but a \( (c_{\text{key}} \times c_{\text{val}}) \) memory\(^1\) as similar channel attention module.

Template matching: Template matching is one of the traditional methods in the tracking task. It generates a template and calculates similarity with input as a matching operation. Most works match a feature map from a given image and a template following the siamese network [1], but A-GAME [11] designed a target distribution by a mixture of Gaussian in an embedding space. It predicted posterior class probabilities for matching. RANet [40] applied a ranking system to the matching process between multiple templates and input for extracting reliable results. FEELVOS [33] calculated distance map by local and global matching for better robustness. SiamMask [37] used a depth-wise operation for fast matching and makes a template from a bounding box annotation without accurate annotated mask of a target.

Consistency Loss: Consistency loss is widely used for improving performance in semi-supervised learning, enhance robustness from perturbation to input, enable stable training under specific constraints, and so on [10, 21, 46]. In VOS, consistency usually means temporal coherence between neighboring frames by additional clue from optical flow. [32, 35, 41].

3 METHOD

In this section, we present our semi-VOS model. Section 3.1 introduces the whole model architecture and how to manage multi-object VOS. Section 3.2 explains the details of template attention module for long-term matching. We also describe how to update the template and how to produce a similarity map. Finally, Section 3.3 demonstrates our temporal consistency loss and how to define new ground truth for mitigating error propagation between neighboring frames.

3.1 Overall TTVOS Architecture

We propose a new architecture for VOS as shown in Fig. 2. Our TTVOS consists of feature extraction, template matching, decoding, and template update stages. The template matching is composed of a short-term matching and a long-term matching. The short-term matching enhances localization property by using previous information. This uses a small feature map for producing a coarse segmentation map. However, this incurs two problems: 1) Utilizing only the information of the previous frame causes the output masks overly dependent on previous results. 2) This can not handle shape-changing nor manifest detailed target shape due to a small feature map. To resolve these problems, we propose long-term matching as an adaptive template matching method. This template is initialized from the given first frame condition and updated at each frame. Therefore, it can consider the whole frames and track gradually changing objects. This module uses a larger feature map for getting more detailed information for generating accurate masks. After then, our model executes decoding and updates each templates step by step.

A backbone extracts feature maps \( fN_t \) from the current frame, where \( fN_t \) denotes a feature map at frame \( t \) with an \( 1/N \)-sized width and height compared to the input. Short-termmatching uses a small feature map \( f16_t \) and the previous frame information for target localization: \( f16_{t-1} \) is concatenated with a previous mask heatmap \( \hat{H}_{t-1} \), which consists of two channels containing the probability of background and foreground respectively. After then, this concatenated feature map is forwarded by several convolution layers for embedding localization information from the previous frame. This information is blended with \( f16_t \) to get an enhanced localization.

\(^1\)\( h \) and \( w \) are the height and the width of an input feature map for constructing memory, and \( c_{\text{key}} \) and \( c_{\text{val}} \) are the number of channels for the key and value feature maps.
We conjecture that pixels inside a target object have a distinct embedding vector distinguished from non-target object pixels. Our model is designed to find this vector by self-attention while suppressing the irrelevant information of the target object. Each current embedding vector updates a previous long-term template by weighted-average at each frame. After then, the proposed module generates a similarity map by template matching to enhance the detailed region as shown in Fig. 3.

For constructing the current embedding vector, the backbone feature \( f_{8_t} \) and the previous estimated mask heatmap \( \hat{H}_{t-1} \) are concatenated to suppress information far from the target object. In Fig. 3, the concatenated feature map is denoted as \( X'_{t-1} \). \( X'_{t-1} \) is forwarded to two separate branches \( f(\cdot) \) and \( g(\cdot) \), making \( f(X'_{t-1}), g(X'_{t-1}) \in \mathbb{R}^{c_{tp} \times H \times W} \). After then, the feature maps are reshaped to \( c_{tp} \times HW \) and produced to generate an embedding matrix \( I \) as follows:

\[
I = \sigma(f(X'_{t-1}) \times g(X'_{t-1})^T) \in \mathbb{R}^{c_{tp} \times c_{tp}}.
\]

Here, \( \sigma \) is a softmax function applied row-wise. \( I_{i,j} \) is the \((i,j)\)th element of \( I \), corresponds to an \( i \)th channel’s view about a \( j \)th channel information by dot-producting along \( HW \) direction. \( X'_{t-1} \) hampers the inflow of information far from the target object by \( \hat{H}_{t-1} \). Thus \( I_{i,j} \) considers only pixels inside or near the target object, and this operation is similar to global pooling and region-based operation [3] in terms of making one representative value from the whole \( HW \)-sized channel and concentrating on a certain region. For example, if the hexagon in Fig. 3(a) indicates the estimated location of the target from the previous mask, the information outside of the hexagon is suppressed. Then \( f(X'_{t-1}) \) and \( g(X'_{t-1}) \) are compared with each other along the whole \( HW \) plane. If the two channels are similar, the resultant value of \( I \) will be high (red pixel in Fig. 3(a)); otherwise, it will be low (blue pixel). Finally, we have \( c_{tp} \) embedding vectors of size \( 1 \times c_{tp} \) containing information about the target object. The final long-term template \( T_{P_t} \) is updated by weighted-average of the embedding matrix \( I \) and the previous template \( T_{P_{t-1}} \) as below:

\[
T_{P_t} = \frac{t-1}{t} T_{P_{t-1}} + \frac{1}{t} I_t.
\]

The template attention module generates a similarity map \( S_t \in \mathbb{R}^{c_{tp} \times H \times W} \) by attending on each channel of the query feature map \( q(X_t) \in \mathbb{R}^{c_{tp} \times H \times W} \) through the template \( T_{P_t} \) as follows:

\[
S_t = T_{P_t} \times q(X_t).
\]

In doing so, the previous estimated mask heatmap \( \hat{H}_{t-1} \) enhances the backbone feature map \( f_{8_t} \) around the previous target object location by forwarding the concatenated feature to a convolution layer resulting in a feature map \( X_t \). Then, \( X_t \) is forwarded to several convolution layers to generate a query feature map \( q(X_t) \) as shown in Fig. 3. In Eq. (3), the similarity is measured between each row of \( T_{P_t} \) (template vector) and each spatial feature from \( q(X_t) \), both of which are of a length \( c_{tp} \). When the template vector is similar to the spatial feature, the resultant \( S_t \) value will be high (red pixel in Fig. 3(a)). Otherwise, it will be low (blue in Fig. 3(a)). After then,
1. The size of $f(X_{t-1})$ and $g(X_{t-1})$ is $c_{zp} \times H \times W$, but we draw feature maps as $c_{zp} \times H \times W$ for the sake of convenient understanding. (b) The detailed structure of a template attention module and a template update. An operation (a,b,c) denotes the input channel, output channel, and kernel size of convolution operation, respectively.

```
Table 1: The complexity and accuracy comparison between GC and ours when the input image size is 480 × 853. Read, Seg, and updates mean the requirement of FLOPS for reading a memory or a template, making a segmentation mask without a decoding stage, and updating a memory or a template. Our method reduces lots of computations for updating the template.
```

|       | Read | Seg | Update | #Param | J&F |
|-------|------|-----|--------|--------|-----|
| GC    | 1.05 | 36.8| 37.1   | 38     | 86.6|
| Ours  | 0.08 | 5.29| 0.06   | 1.6    | 79.5|

The reason why we use $H_{t-1}$ instead of $H_{t-1}$ is illustrated in Fig. 4. Fig. 4(b) shows ground truth masks, and (c) is the estimated masks at frame $t-1$ (top) and $t$ (bottom). First row of Fig. 4(e) is a visualization of $(H_t - H_{t-1})$ that guides the estimation to maintain the false positive region from the frame $t - 1$ to $t$. Second row of Fig. 4(e) is a visualization of $(H_t - H_{t-1})$ that guides the estimation to remove false positive region of the frame $t - 1$. Fig. 4(d) is marked by blue color for denoting false estimation results comparing between (b) and (c). As shown in Fig. 4(d), the transition matrix $\pi_t$ helps...
reducing the false positive region from frame $t - 1$ to $t$. With $L_{tc}$, the overall loss becomes:

$$Loss = CE(y_t, \hat{y}_t) + \lambda L_{tc},$$

where $\lambda$ is a hyper-parameter that controls the balance between the loss terms, and we set $\lambda = 5$. $CE$ denotes the cross entropy between the pixel-wise ground truth $y_t$ at frame $t$ and its predicted value $\hat{y}_t$.

### 4 EXPERIMENT

Here, we show various evaluations by using DAVIS benchmarks [24, 25]. DAVIS16 is a single object task consisting of 30 training videos and 20 validation videos, while DAVIS17 is a multiple object task with 60 training videos and 30 validation videos. We evaluated our model by using official benchmark code\(^2\). The DAVIS benchmark reports model accuracy by average of mean Faccard index $J$ and mean boundary score $F$. $J$ index measures overall accuracy by comparing estimated mask and ground truth mask. $F$ score focuses more contour accuracy by delimiting the spatial extent of the mask.

**Implementation Detail:** We used HRNetV2-W18-Small-v1 [36] for a lightweight backbone network and initialized it from the pre-trained parameters from the official code\(^3\). We froze every backbone layer except the last block. The size of the smallest feature map is 1/32 of the input image. We upsampled the feature map and concatenated it with the second smallest feature map whose size is 1/16 of the input image. We used ADAM optimizer for training our model. First, we pre-trained with synthetic video clip from image dataset, after then we trained with video dataset with single GPU following [11, 22, 33, 37].

**Pre-train with images:** We followed [16, 22, 40] pre-training method, which applies random affine transformation to a static image for generating synthetic video clip. We used the saliency detection dataset MSRA10K [5], ECSSD [43], and HKU-IS [15] for various static images. Synthetic video clips consisting of three frames with a size of $240 \times 432$ were generated. We trained 100 epochs with an initial learning rate to $1e^{-4}$ and a batch size to 24.

**Main-train with videos:** We initialized the whole network with the best parameters from the previous step and trained the model to video dataset. We used a two-stage training method; for the first 100 epochs, we only used Youtube-VOS with $240 \times 432$ image. We then trained on the DAVIS16 dataset with $480 \times 864$ image for an additional 100 epochs. Both training, we used 8 consecutive frames with a batch size to 8 and set an initial learning rate to $1e^{-4}$.

### 4.1 DAVIS Benchmark Result

**Comparison to state-of-the-art:** We compared our method with other recent models as shown in Table 2. We report backbone models and training datasets for clarification because each model has a different setting. Furthermore, we also show additional results with ResNet50 because some recent models utilized ResNet50 for extracting features.

Our result shows the best accuracy among models with similar speed. Specifically, SiamMask is one of the popular fast template matching methods, and our model has better accuracy and speed than SiamMask on both DAVIS16 and DAVIS17 benchmark. When we used ResNet50, our model has better or competitive results with FRTM-VOS, A-GAME, RANet, and FEELVOS. Also, this ResNet50 based model decreases DAVIS16 accuracy by 2.8% but the speed becomes 1.6 times faster than GC. Therefore, our method achieves favorable performance among fast VOS models and reduces the performance gap from the online-learning and memory network based models.

**Ablation Study:** For proving our proposed methods, we performed an ablative analysis on DAVIS16 and DAVIS17 benchmark as shown in Table 3. SM and LM mean short-term matching and long-term matching, respectively. When we do not use short-term matching or long-term matching, we replaced the original matching method into concatenating the previous mask heatmap and the current feature map. After then the concatenated feature map is forwarded by several convolution layers. $L_{tc}$ denotes using temporal consistency loss. Without this, the model only uses a cross entropy loss. $M$ denotes using the original ground truth mask for the initial condition; if $M$ is not checked, a box-shaped mask is used for the initial condition like SiamMask. $Exp1$ is using only short-term matching, and $Exp2$ is using only long-term matching. $Exp3-6$ uses both matching methods. Table 3 is the corresponding accuracy for each ablation experiment, and Fig. 6 visualizes efficacy of each ablation experiment.

---

2. [https://github.com/HRNet/HRNet-Semantic-Segmentation](https://github.com/HRNet/HRNet-Semantic-Segmentation)

3. [https://github.com/davisvideochallenge/davis2017-evaluation](https://github.com/davisvideochallenge/davis2017-evaluation)

---

**Figure 4:** (a)-(d) frame $t-1$ and $t$ from top to bottom. (a) Input image. (b) Ground truth. (c) Our result. (d) Estimated mask with color marking. Blue color means wrong segmentation result, and the blue region in frame $t$ is corrected from frame $t-1$. (e) Visualizing $\pi t$. Top: $H_t - H_{t-1}$, Bottom: $H_t - H_t - H_{t-1}$. $H_t - H_{t-1}$ cannot remove false positive region in the top of (e).
### Table 2: Quantitative comparison on DAVIS benchmark validation set.
OL and Memory denotes online-learning approach and memory network approach. YTB is using Youtube-VOS for training. Seg is segmentation dataset for pre-training by Pascal [8] or COCO [18]. Synth is using saliency dataset for making synthetic video clip by affine transformation.

| Method          | Backbone | Model Method | Train Dataset | DV17 | DV16 | FPS |
|-----------------|----------|--------------|---------------|------|------|-----|
| OnAVOS [34]     | VGG16    | o -          | - o           | 67.9 | 85.5 | 0.08|
| OSVOS-S [20]    | VGG16    | o -          | - o           | 68.0 | 86.5 | 0.22|
| FRTM-VOS [26]   | ResNet101| o o          | - o           | 76.7 | 83.5 | 21.9|
| STM [22]        | ResNet50 | - o          | o - o         | 81.8 | 89.3 | 6.25|
| GC [16]         | ResNet50 | - o          | o - o         | 71.4 | 86.6 | 25.0|
| OSMN [44]       | VGG16    | - -          | - o           | 54.8 | 73.5 | 7.69|
| RANet [40]      | ResNet101| - -          | - - o         | 65.7 | 85.5 | 30.3|
| A-GAME [11]     | ResNet101| - -          | o - o         | 70   | 82.1 | 14.3|
| FEELVOS [33]    | Xception 65 | - -       | o o o -         | 71.5 | 81.7 | 2.22|
| SiamMask [37]   | ResNet50 | - -          | o o o -        | 56.4 | 69.8 | 55.0|
| TTVOS (Ours)    | HRNet    | - -          | o - o          | 58.7 | 79.5 | 73.8|
| TTVOS-RN (Ours) | ResNet50 | - -          | o - o          | 67.8 | 83.8 | 39.6|

### Table 3: Ablation study on DAVIS16 and DAVIS17. SM, LM, TC means short-term matching, long-term matching and temporal consistency loss. Lup represents updating long-term template at every frame, and M is using original ground truth mask for initial condition.

| Exp | SM | LM | Lup | TC | M | DV17 | DV16 |
|-----|----|----|-----|----|---|------|------|
| 1   | o  | -  | -   | -  | o | 57.0 | 75.9 |
| 2   | -  | o  | o   | -  | o | 54.5 | 78.8 |
| 3   | o  | o  | o   | -  | o | 57.5 | 77.1 |
| 4   | o  | o  | o   | o  | o | 58.6 | 77.6 |
| 5   | o  | o  | -   | o  | o | 57.2 | 77.4 |
| 6   | o  | o  | o   | o  | o | 58.7 | 79.5 |

Figure 5: Example of parkour for frame 1, 34 and 84 from top to Bottom. Column (a) shows input images overlapped with the ground truth masks. RM-LongM denotes estimated results removing long-term matching information by replacing to zeros.

Figure 6: Horsejump-high example of ablation study for frame 3 and 37 from top to bottom. (a) Ground truth. (b) Using only short-term matching. (c) Using only long-term matching. (d) Our proposed method (Exp6).

We found that short-term matching helps maintain objects ID from localization clue, and long-term matching improves mask quality by enhancing the detailed regions. For example, Exp1 keeps object ID but fails to make an accurate mask for horse legs, as shown in Fig. 6(b). On the contrary, Exp2 makes accurate shape but loses green-object (rider) ID as shown in Fig. 6(c). Exp2 shows performance degradation on multi-object tracking task (DAVIS 17) due...
to failure in maintaining object ID, even it generates more accurate masks than Exp1. Therefore, Exp1 achieves better performance in DAVIS17, and Exp2 shows high accuracy in DAVIS16. Exp3 gets every advantage from both template matching methods, and Fig. 6(d) is our proposed method results (Exp6), which do not lose object ID and generate delicate masks with high performance on both benchmarks.

Exp4-6 explain why our model shows better performance than SiamMask, even using a more lightweight backbone. The initial condition of the box shape mask does not degrade performance a lot comparing with Exp6. However, when the model does not update the long-term template, the accuracy degrades a lot from our proposed method.

**Temporal Consistency Loss**: We conducted further experiments for proving the efficacy of our temporal consistency loss with FRTM-VOS, which is one of the fast online-learning methods, using ResNet101 and ResNet18 for the backbone network. We implemented our proposed loss function based on FRTM-VOS official code\(^1\), and followed their training strategy. Our proposed loss is more useful in the lightweight backbone network (ResNet18) as shown in Table 4. When we applied our loss to the ResNet101 model, the accuracy on DAVIS17 decreased slightly by 0.1%, but it increased 1.7% on DAVIS16. In the ResNet18 model, we improved the accuracy a lot on both DAVIS17 and DAVIS16. We conjecture that using our loss not only improves mask quality but also resolves a problem of overfitting due to fine-tuning by a given condition.

| Backbone | DAVIS17 | DAVIS16 |
|----------|---------|---------|
| FRTM-VOS [26] | ResNet101 | 87.6 | 83.5 |
|           | ResNet18 | 70.2 | 78.5 |
| with TC Loss | ResNet101 | 76.6 | 85.2 |
|           | ResNet18 | 71.8 | 82.0 |

Table 4: DAVIS17 and DAVIS16 results when additional applying temporal consistency loss (TC Loss).

**REFERENCES**

\(1\) Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In **ECCV 2016 Workshops**, 850–865.

\(2\) S. Caelles, K.K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. 2017. One-Shot Video Object Segmentation. In Computer Vision and Pattern Recognition (CVPR).

\(3\) Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2016. Region-based semantic segmentation with end-to-end training. In European Conference on Computer Vision. Springer, 381–397.

\(4\) Jingdong Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. 2017. SegFlow: Joint learning for video object segmentation and optical flow. In Proceedings of the IEEE international conference on computer vision. 686–695.

\(5\) Ming-Ming Cheng, Nila J Mitra, Xiaolei Huang, Philip HS Torr, and Shin-Min Hu. 2014. Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence 37, 3 (2014), 569–582.

\(6\) Suhwan Cho, MyeongAh Cho, Tae-young Chang, Heunsang Lee, and Sangyoun Lee. 2020. CRVOS: Clue Refining Network for Video Object Segmentation. arXiv preprint arXiv:2002.03511 (2020).

\(7\) Suyog Dutta Jain, Bo Xiong, and Kristen Grauman. 2017. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3664–3673.

\(8\) Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. International journal of computer vision 111, 1 (2015), 98–136.

\(9\) Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yu-Peng Tan. 2018. Motion-guided cascaded refinement network for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1400–1409.

\(10\) Junsu Jeong, Seongeui Lee, Jeesoo Kim, and Nojun Kwak. 2019. Consistency-based semi-supervised learning for object detection. In Advances in neural information processing systems. 10759–10768.

\(11\) Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shakhbaz Khan, and Michael Felsberg. 2019. A generative appearance model for end-to-end video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8953–8962.

\(12\) Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. 2017. Lucid data dreaming for object tracking. In The DAVIS Challenge on Video Object Segmentation.

\(13\) Junyeong Kim, Mimik Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. 2019. Progressive attention memory network for movie story question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8341–8346.

\(14\) Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. 2004. Online learning with kernels. IEEE transactions on signal processing 52, 8 (2004), 2165–2176.

\(15\) Guanbin Li and Yizhou Zhang. 2017. Learning video object segmentation from static images. In Proceedings of the IEEE conference on computer vision and pattern recognition. 5455–5463.

\(16\) Yu Li, Zhaozhi Shen, and Ying Shan. 2020. Fast Video Object Segmentation using the Global Context Module. In The European Conference on Computer Vision (ECCV).

\(17\) Fanqing Lin, Tao Chou, and Tony Martinez. 2020. Flow Adaptive Video Object Segmentation. Image and Vision Computing 84 (2020), 103864.

\(18\) Tsung-Yi Lin, Michael Maire, Serguei Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision (ECCV). 116–131.

\(19\) K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. 2018. Video object segmentation without temporal information. IEEE transactions on pattern analysis and machine intelligence 41, 6 (2018), 1515–1530.

\(20\) Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence 41, 8 (2018), 1979–1993.

\(21\) Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Woo Kim. 2019. Video object segmentation using space-time memory networks. In Proceedings of the IEEE International Conference on Computer Vision. 9226–9235.

\(22\) Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bern Schiele, and Alexander Sorkine-Hornung. 2017. Learning video object segmentation from static images. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2663–2672.

\(23\) Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 724–732.
[25] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. arXiv:1704.00673 (2017).

[26] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shabbaz Khan, and Michael Felsberg. 2020. Learning Fast and Robust Target Models for Video Object Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7406–7415.

[27] Doyen Sahoo, Quang Pham, Jing Lu, and Steven C. H. Hoi. 2018. Online Deep Learning: Learning Deep Neural Networks on the Fly. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 2660–2666. https://doi.org/10.24963/ijcai.2018/369

[28] Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J Black. 2016. Optical flow with semantic segmentation and localized layers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3889–3898.

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1874–1883.

[30] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In Advances in neural information processing systems. 2440–2448.

[31] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J. Black. 2016. Video segmentation via object flow. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3899–3908.

[32] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J. Black. 2016. Video Segmentation via Object Flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

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

[34] Paul Voigtlaender and Bastian Leibe. 2017. Online Adaptation of Convolutional Neural Networks for Video Object Segmentation. In British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017. BMVA Press.

[35] Sebastian Volz, Andres Bruhn, Levi Valgaerts, and Henning Zimmer. 2011. Modeling temporal coherence for optical flow. In 2011 International Conference on Computer Vision. IEEE, 1116–1123.

[36] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2019. Deep High-Resolution Representation Learning for Visual Recognition. TPAMI (2019).

[37] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1328–1338.

[38] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. 2018. Semi-supervised video object segmentation with super-trajectories. IEEE transactions on pattern analysis and machine intelligence 41, 4 (2018), 985–998.

[39] Xiaodong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7794–7803.

[40] ZiQun Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. 2019. Ranet: Ranking attention network for fast video object segmentation. In Proceedings of the IEEE international conference on computer vision. 3978–3987.

[41] Joachim Weickert and Christoph Schnörr. 2001. Variational optic flow computation with a spatio-temporal smoothness constraint. Journal of mathematical imaging and vision 14. 3 (2001), 245–255.

[42] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. arXiv preprint arXiv:1410.3916 (2014).

[43] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical saliency detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1155–1162.

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

[45] Guanyu Zhou, Kihyuk Sohn, and Honglak Lee. 2012. Online incremental feature learning with denoising autoencoders. In Artificial intelligence and statistics. 1453–1461.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision. 2223–2232.

[47] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. 2019. Asymmetric normalization in deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7406–7415.

[48] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1328–1338.