Reliable Propagation-Correction Modulation for Video Object Segmentation

Xiaohao Xu, [1,2*] **Jinglu Wang,** [2] **Xiao Li,**[2] **Yan Lu** [2]

[1] Huangzhong University of Science and Technology
[2] Microsoft Research Asia
xxh11102019@outlook.com, {jinglwa, xili11, yanlu}@microsoft.com

Abstract

Error propagation is a general but crucial problem in online semi-supervised video object segmentation. We aim to suppress error propagation through a correction mechanism with high reliability. The key insight is to disentangle the correction from the conventional mask propagation process with reliable cues. We introduce two modulators, *propagation* and *correction modulators*, to separately perform channel-wise re-calibration on the target frame embeddings according to local temporal correlations and reliable references respectively. Specifically, we assemble the modulators with a cascaded *propagation-correction* scheme. This avoids overriding the effects of the reliable correction modulator by the propagation modulator. Although the reference frame with the ground truth label provides reliable cues, it could be very different from the target frame and introduce uncertain or incomplete correlations. We augment the reference cues by supplementing reliable feature patches to a maintained pool, thus offering more comprehensive and expressive object representations to the modulators. In addition, a reliability filter is designed to retrieve reliable patches and pass them in subsequent frames. Our model achieves state-of-the-art performance on YouTube-VOS18/19 and DAVIS17-Val/Test benchmarks. Extensive experiments demonstrate that the correction mechanism provides considerable performance gain by fully utilizing reliable guidance.

Introduction

Semi-supervised video object segmentation (VOS), also known as mask tracking, aims at segmenting target objects in a video sequence given the ground truth mask at the first (or reference) frame. Recently, sequence-to-sequence methods (Vaswani et al. 2017; Duke et al. 2021) achieve impressive results but suffer from relatively high cost. Online methods (Perazzi et al. 2017; Oh et al. 2018; Wang et al. 2018, 2019), taking only the current frame with image-wise references as input, are more practical for fast and streaming applications. We focus on online methods in this paper.

The Semi-supervised VOS problem is usually formulated as a maximum a posterior (MAP) problem, conditioning on the target frame, preceding and reference frames and labels.
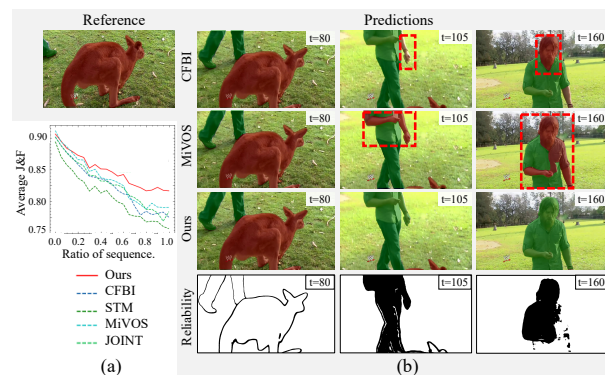

Figure 1: Our model can suppress error propagation in VOS with a reliable propagation-correction mechanism. (a) Average overall performance (J&F) on YouTube-VOS 19 validation set over time. Ours has the least performance decay. (b) Considering the large appearance transition from the reference to targets, our model achieves much better results compared to CFBI (Yang, Wei, and Yang 2020) and MiVOS (Cheng, Tai, and Tang 2021a) (red rectangles bound error regions). The binary reliability maps indicate reliable (white) and uncertain (black) patches.

Considering the probabilistic model of online VOS, the current label can be predicted from a frame-by-frame propagation path or a direct translation path from the reliable reference label. To exploit local temporal consistency, many methods (Oh et al. 2018; Perazzi et al. 2017; Hu, Huang, and Schwing 2017; Voigtlaender and Leibe 2017; Cheng et al. 2017, 2018; Hu, Huang, and Schwing 2017) follow the propagation path to perform mask propagation, but errors may accumulate over time due to the inevitable prediction uncertainty in each iteration. The reference frame with a ground truth label provides reliable object cues, thus having the potential to suppress error propagation (Li et al. 2020). Recent methods (Voigtlaender et al. 2019; Yang, Wei, and Yang 2020) demonstrate that even naively manipulating references by feature concatenation and matching could improve the VOS performance. This encourages us to make full use of reliable reference cues to correct errors during the mask propagation. However, the target frame may turn out to

be very different from the reference when time goes by, losing explicit correspondences to the reference. For example, in Fig. 1 (b), the reference only containing part of the object is not comprehensive to represent the whole object, *i.e.*, the foot in reference. In this case, estimating correlations between the reference and target is uncertain and incomplete, which may lead to negative impacts on the VOS task.

Network modulation, which recalibrates feature embeddings with additional conditions, has achieved great success in VOS (Yang et al. 2018; Yang, Wei, and Yang 2020). The modulation operation is light-weight and can be performed per frame, which fulfils the streaming requirement. The key to modulation is to construct expressive conditional weights and extract highly correlated embeddings. For the VOS task, modulation weights need to be representative for reference objects and embeddings also need to encode reliable correlations between the reference and the target. In this paper, we propose a new end-to-end framework for VOS with reliable propagation-correction modulation, which can provide representative object proxies weights for modulation and consolidate target embeddings in a cascaded assembly of propagation and correction modulators.

To perform correction with high reliability, we augment the translation path from the reference to a more comprehensive correction path. Since object cues in the reference frame may be incomplete, we progressively supplement them with reliable information in each iteration. A reliable patch memory pool is maintained to store the historical reliable feature patches, which is further utilized in subsequent frames. The reliable patch pool is for two usages, *i.e.*, augmenting the object proxy with comprehensive information to obtain expressive modulation weights and consolidating frame embedding with more reliable correlations. As for the network design, we introduce two types of modulator, *i.e.*, propagation and correction modulator, which separately augment embedding according to the local temporal correlation cues in the propagation path and the reliable reference cues from the correction path. To avoid overriding the correction effect, the correction modulator is inserted after the propagation modulator. We also propose a reliability filter to assess the prediction quality. From the example reliability map in Fig. 1 (b), regions with large appearance change from reference are predicted as uncertain while other reliable regions can be passed to the following frames for correction. Our experiments demonstrate that the assembly of propagation and correction modulators has a considerable impact on VOS performance. Our contributions are three-fold.

- We propose a new reliable propagation-correction modulation model for VOS, which significantly suppresses error propagation (see precision decay curve in Fig. 1 (a)). Our model achieves the state-of-the-art performance on both YouTube-VOS and DAVIS17 benchmarks.
- We disentangle the reliable correction from the conventional erroneous mask propagation process with separate memory modulators.
- We augment both the object proxy and target embedding with comprehensive reliable cues to reinforce the correction modulation.

Related Work

Propagation-based VOS. Propagation-based methods utilize the semantic or spatial cues from the previous frame to predict the mask of the current frame. Early methods (Perazzi et al. 2017; Caelles et al. 2017; Hu, Huang, and Schwing 2017; Khoreva et al. 2019) utilize online-learning method to eliminate the drifting problem but is time-consuming. Optical flow (Tsai, Yang, and Black 2016; Hu, Huang, and Schwing 2017; Cheng et al. 2017; Xu et al. 2018b) and object tracking also prove to be useful guidance for mask propagation. Although propagation-based models can secure good temporal consistency(Caelles et al. 2017), these propagation-based methods are prone to error accumulation, which may largely degrade the VOS performance especially for long video clips(Liang et al. 2020b).

Matching-based VOS. Matching-based methods learn an embedding space for target objects. (Chen et al. 2018b; Hu, Huang, and Schwing 2018; Zeng et al. 2019) directly construct the correspondence between the current frame with the first frame. (Lin, Qi, and Jia 2019; Voigtlaender et al. 2019; Wang et al. 2019; Yang, Wei, and Yang 2020) further leverage both the first and the previous frames. Several recent methods (Hu, Huang, and Schwing 2018; Liang et al. 2020a; Duke et al. 2021) turn to use several latest frames to further improve the local temporal guidance. Moreover, STM-based networks (Oh et al. 2019; Seong, Hyun, and Kim 2020; Lu et al. 2020; Liang et al. 2020b,c; Cheng, Tai, and Tang 2021b; Wang et al. 2021; Xie et al. 2021; Hu et al. 2021; Seong et al. 2021) boost the performance with memory networks that memorize information from past frames for further reuse, which relieve the error propagation to some extent. However, how to reduce uncertainty propagation is still a hard problem and hasn't been tackled perfectly. Our method disentangles the guidance by reliability and further suppresses uncertainty with a carefully designed scheme.

Conditional modulation. Recently, (Yang et al. 2018; Yang, Wei, and Yang 2020; Li et al. 2021) introduce conditional modulation methods to tackle video-related tasks like video object segmentation for instance-based, or spatial guidance. However, the modulation weights or embeddings for mask propagation in the video are inevitably volunteered to error. To collaboratively suppress error propagation and make full use of the power of conditional modulation, we propose a new conditional modulation method called *reliable conditional modulation* to tackle VOS.

Preliminaries

We first review the probabilistic model of frame-to-frame VOS and analyze it from two aspects, *i.e.*, the frame-by-frame propagation path and the correction path from reliable reference. We also introduce a widely used measurement for prediction reliability.

Probabilistic Model of Frame-to-frame VOS

Given all available observations $\mathbf{x}_{1:t}$ up to t-th frame, the label up to t-th frame $\mathbf{y}_{1:t}$ is predicted by maximum a posterior (MAP) estimation:
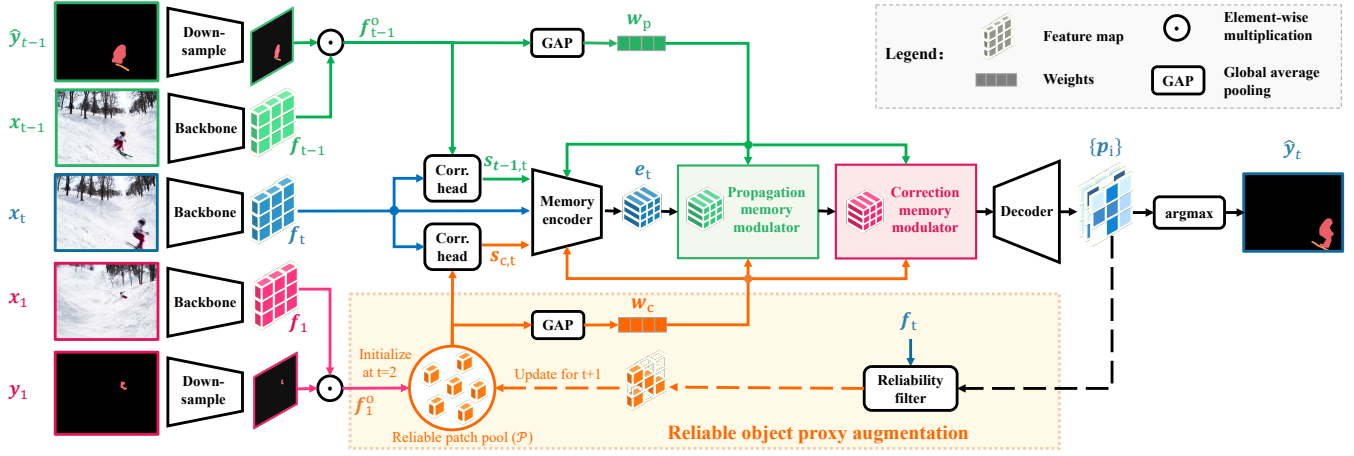
Figure 2: Overview of the proposed framework. We disentangle the correction mechanism from the frame-to-frame mask propagation process. We assemble a cascaded scheme of propagation-correction modulators to leverage local temporal correlations and reliable references in order. We also augment the reference cues by supplementing reliable feature patches to a maintained pool, thus offering more comprehensive and expressive object representations to the modulators. A reliability filter is introduced to filter out uncertain patches for subsequent frames.

$$p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) = \frac{p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})p(\mathbf{y}_t)}{p(\mathbf{x}_{1:t})} \propto p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})p(\mathbf{y}_t) \quad (1)$$

Here $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$ is the observation model, which is usually estimated by the likelihood $p(\mathbf{x}_{1:t}|\mathcal{D})$, where $\mathcal{D}$ denotes the training data. $p(\mathbf{y}_{1:t-1}|\mathbf{x}_{1:t-1})$ is the posterior up to previous frame. $p(\mathbf{y}_{1:t})$ is the prior model and could be unfolded with the first-order markov assumptions:

$$p(\mathbf{y}_{1:t}) = p(\mathbf{y}_t|\mathbf{y}_{t-1})p(\mathbf{y}_{1:t-1}) = p(\mathbf{y}_1)\Pi_2^t p(\mathbf{y}_i|\mathbf{y}_{i-1}) \quad (2)$$

We then instantiate Equation 1 with the propagation and correction path respectively.

Propagation path. We assume that the observation model is conditionally independent given the states, *i.e.*, $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t}) = \Pi_1^t p(\mathbf{x}_i|\mathbf{y}_i)$. The posterior takes the form

$$\begin{aligned} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) &\propto \Pi_1^t p(\mathbf{x}_i|\mathbf{y}_i)\Pi_2^t p(\mathbf{y}_i|\mathbf{y}_{i-1})p(\mathbf{y}_1) \\ &\propto \Pi_1^t p(\mathbf{x}_i|\mathbf{y}_i)\Pi_2^t p(\mathbf{y}_i|\mathbf{y}_{i-1}) \end{aligned} \quad (3)$$

Note $p(\mathbf{y}_1)$ is omitted since the label of the first frame $\mathbf{y}_1$ is given. Therefore, we observe that prediction uncertainty of $p(\mathbf{y}_t|\mathbf{y}_{t-1})$ will accumulate over time, which lead to error propagation.

Correction path. We first consider the direct translation from the reliable reference frame $\mathbf{x}_1$ to the t-th frame $\mathbf{x}_t$ for correction. The posterior takes the form

$$\begin{aligned} p(\mathbf{y}_1, \mathbf{y}_t|\mathbf{x}_1, \mathbf{x}_t) &= \frac{p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{y}_1, \mathbf{x}_1)p(\mathbf{x}_1, \mathbf{y}_1)}{p(\mathbf{x}_1, \mathbf{x}_t)} \\ &\propto p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{y}_1, \mathbf{x}_1) \end{aligned} \quad (4)$$

Again, $p(\mathbf{x}_1, \mathbf{y}_1)$ is omitted since $\mathbf{x}_1$ and $\mathbf{y}_1$ are given. The joint condition probability $p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{y}_1, \mathbf{x}_1)$ corresponds to the joint similarity of observations and labels between target and reference frames. Since labels represent object masks,

the joint similarity can be considered as an object-aware similarity. Thus, prediction in the correction path is highly correlated to the object-aware similarity between target and reference frames. Since reference in the first frame may not be comprehensive, we will augment it during the iterations.

Prediction reliability. Prediction uncertainty of deep neural networks is difficult to estimate accurately, while it is highly correlated with information entropy (Feder and Merhav 1994). Here, we employ the Shannon entropy to measure the reliability of prediction in each iteration:

$$H(I) = -\sum_{i=1}^{N+1} P(\frac{e^{\mathbf{P}i}}{\sum_{j=1}^{N+1} e^{\mathbf{P}j}}) \log P(\frac{e^{\mathbf{P}i}}{\sum_{j=1}^{N+1} e^{\mathbf{P}j}}) \quad (5)$$

where $\mathbf{p}_i, i \in \{1, ..., N+1\}$ indicate the probability maps of N objects and background.

Reliable Propagation-Correction Modulation

In this section, we first describe the overview of our pipeline and then elaborate on the reliable modulation mechanism. Finally, the network design is detailed.

Pipeline Overview

Fig. 2 illustrates the overview of our framework. Our goal is to predict the object mask $\hat{\mathbf{y}}_t$ of the target frame $\mathbf{x}_t$, given the first frame $\mathbf{x}_1$ and its object mask $\mathbf{y}_1$, as well as the previous frame $\mathbf{x}_{t-1}$ and predicted object mask $\hat{\mathbf{y}}_{t-1}$.

We first extract features maps $\mathbf{f}_1, \mathbf{f}_{t-1}, \mathbf{f}_t$ of frames $\mathbf{x}_1, \mathbf{x}_{t-1}, \mathbf{x}_t$ from a shared backbone respectively. To make features object-aware, we mask the features with corresponding object masks, $f_{t'}^o = \mathbf{f}_{t'} \odot \mathbf{y}_{t'}, t' = \{1, t-1\}$. For the propagation path, we define an object proxy w_p representing a image-level object feature by applying a global average pooling (GAP) on f_{t-1}^o. Object-aware inter-frame correlation for propagation path is then represented with a similarity map $\mathbf{s}_{t-1,t}$ between the target feature $\mathbf{f_t}$ and masked
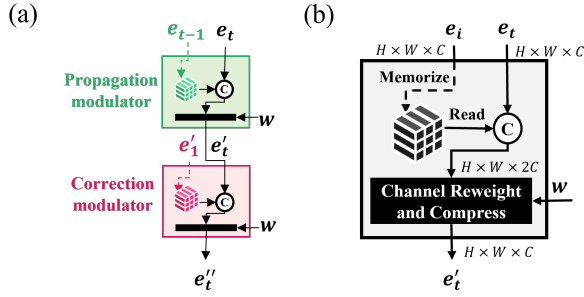
Figure 3: (a) Cascaded propagation-correction ($P2C$) modulator scheme. (b) Basic modulator block.

previous feature f_{t-1}^o. For the correction path, we also define an object proxy w_c and a similarity map $\mathbf{s}_{c,t}$; instead of directly computing them from a reference frame feature, we maintain a reliable feature patch pool containing reliable object feature patches from historical frames. The object proxy and correlation cues from both paths are then further encoded with a memory encoder to form a compact embedding $\mathbf{e}_t$ for the target frame. Given two object proxies w_p and w_c, the current embedding is modulated with the cascaded propagation and correction modulator respectively. The final probability $\mathbf{p}$ for each object is decoded from the modulated embedding with a decoder. In turn, we provide the reliability map $\mathbf{r}_t$ for updating the reliable patch pool $\mathcal{P}$ for the next frame with the reliability filter.

Object Proxy for Correction Path

Reliable object patch pool. For correction path, the object semantics in the first frame is often incomplete. Specifically, the mask may cover only a part of an object. This will highly degrade the correction effects from the reference frame. Thus, we need to augment the object proxy by supplementing new features in the process of mask propagation. However, previous study (Wu et al. 2020) has shown that mask propagation with multiple historical frames may be vulnerable to the influence of inaccurate information, leading to error propagation. To augment object proxy with historical cues while suppressing error propagation, we introduce a *reliable object patch pool* for updating useful feature patches for object representation augmentation and a *reliability filter* to filter out uncertain feature patches.

Specifically, we use a reliability map $\mathbf{r}$ to filter out uncertain patches, i.e., $\mathbf{f_c} = \mathbf{f} \odot \mathbf{y} \odot \mathbf{r}$. During mask propagation, we only recall the highly reliable patches of the object and supplement them to the reliable object patch pool $\mathcal{P}$, thus augmenting the semantic cues of target objects. Details are described in Algorithm 1.

Reliable object proxy. Given the reliable object patch pool $\mathcal{P}$ containing a set of highly reliable patches, we construct the reliable object proxy w_c by apply GAP on all elements in $\mathcal{P}$, i.e., $w_c = \mathrm{GAP}(\{\mathbf{f} \odot \mathbf{y} \odot \mathbf{r}\}, \mathbf{f} \in \mathcal{P})$.

Algorithm 1: Reliable object proxy augmentation

Input: Embedding of current frame $\mathbf{f}_t$, embedding $\mathbf{f}_{t-1}$, mask $\hat{\mathbf{y}}_{t-1}$ and reliability map $\mathbf{r}_{t-1}$ of previous frame, reliable object patch pool $\mathcal{P}$ and its updating time interval τ, timestamp $t \in \{2, ...\}$.
Output: Augmented reliable object proxy w_c and similarity map from the correction path $s_{c,t}$

1: **if t == 2 then**
2:     Let $\mathbf{r}_1 = \mathbf{I}$, $\mathcal{P} = \varnothing$.
3: **end if**
4: **if** $((t-2) \mod \tau) == 0$ **then**
5:     $f_{t-1}^o = \mathbf{f}_{t-1} \odot \hat{\mathbf{y}}_{t-1} \odot \mathbf{r}_{t-1}$
6:     $\mathcal{P} = \mathcal{P} \bigcup f_{t-1}^o$.
7: **end if**
8: $w_c = \frac{\sum_{f \in \mathcal{P}} f}{|\mathcal{P}|}$
9: $s_{c,t} = \max \mathrm{S}(\mathcal{P}, \mathbf{f}_t)$,
    where S is a similarity measure.
10: **return** $w_c, s_{c,t}$.

Propagation-Correction Modulation

While most methods (Oh et al. 2019; Seong, Hyun, and Kim 2020; Lu et al. 2020) employ reference and previous cues equivalently, they do not distinguish their influence on the prediction of the current frame. We address that mask propagation along the propagation path can preserve local temporal consistency, while the reference provides more reliable information. The reliable reference is more suitable to perform a correction role after propagation, and thus such two kinds of cues should be handle in a disentangled way to be made full use of.

We maintain two types of external memory modules to selectively memorize information from the propagation and correction paths, namely, *propagation memory* storing the previous embedding $\mathbf{e}_{t-1}$ and updating at each frame, and *correction memory* storing the embedding of reference frame $\mathbf{e}_1$. Accordingly, we build two modulator ϕ_{prop} and ϕ_{corr} to modulate the current embedding with corresponding object proxy w_p and w_c. Intuitively, we set the correction modulator after the propagation memory as illustrated in Fig. 3 (a) because correlation is reliable and should not be overridden by the propagation modulator.

Our key observation is that the order of modulators in cascaded schemes matters since the uncertainty also relies on the depth of relevant layers in the neural network (Goldfeld et al. 2019). We verified this observation with detailed analysis in the experiment section and confirmed a cascade with propagation-correction order (i.e. $P2C$) outperforms other cascade variants as well as parallel approaches.

Network Design

We detail the implementation of each network module.

Correlation head. The correlation head for calculating similarities between features consists of a set of linear operations in local windows as (Voigtlaender et al. 2019), and then the similarity maps are concatenated with the previous mask and projected into a high-dimensional space. Note
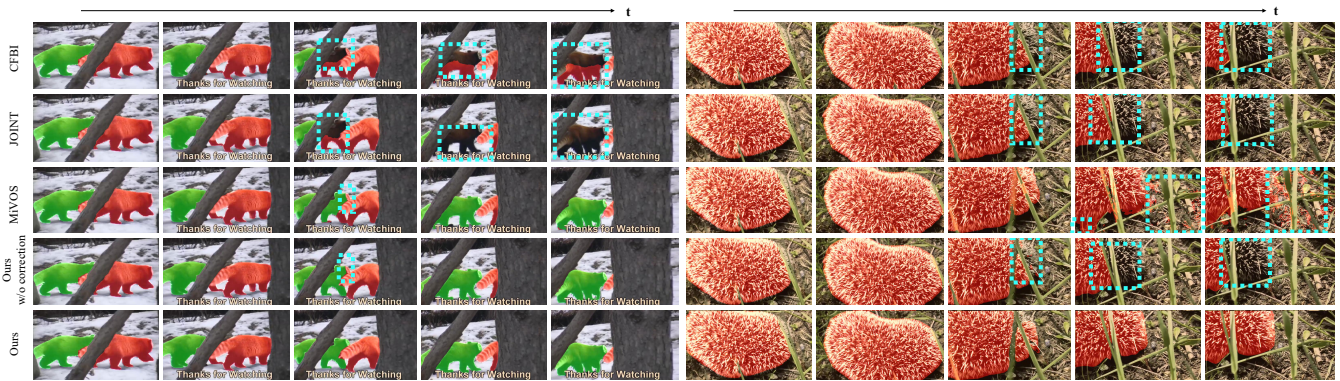
Figure 4: Qualitative comparison to several state-of-the-art methods, CFBI (Yang, Wei, and Yang 2020), JOINT (Mao et al. 2021), MiVOS (Cheng, Tai, and Tang 2021a) on YouTube-VOS 19 validation set. With the reliable correction mechanism, our model can reduce the error regions in the mask propagation process. Error regions are highlighted with blue bounding boxes.

that we follow (Yang, Wei, and Yang 2020) to split features into the foreground and background-masked ones according to corresponding masks and concatenated them together for further processing.

Memory encoder. The memory encoder ϕ_{me} maps the concatenated features $[\mathbf{f}_t, \mathbf{s}_{c,t}, \mathbf{s}_{t-1,t}]$ to a lower dimension and compact embedding $\mathbf{e}_t$. Meanwhile, it consists of a 1×1 convolution layer followed by a series of channel reweighting operations. $\mathbf{e}_t = \phi_{me}([\mathbf{f}_t; \mathbf{s}_{c,t}; \mathbf{s}_{t-1,t}], [w_p; w_c])$, where $[;]$ denotes concatenation.

Modulator block. Fig. 3 (b) illustrates the structure of a basic memory modulator block. A memory modulator block maintains a memory buffer $\mathbf{e}_i \in \mathbb{R}^{H \times W \times C}$ for later usage. During each forward, the memory modulator block inputs memory embedding features $\mathbf{e}_t$ from the current frame, reads out the buffered memory embedding (i.e., $\mathbf{e}_{t-1}$ for propagation modulator or $\mathbf{e}_1$ for correction modulator), concatenates them together and then performs a reweighting operation along channel dimension with w_p or w_c.

Inspired by the theoretical proof (Chen et al. 2011) that low-rank matrix can recover from errors and erasures, we implicitly force the memory embedding to simultaneously keep the low-rank property and encode more useful information by compressing the reweighted embedding to a lower dimension via a 1×1 convolution layer.

Reliability filter. We first compute the Shannon entropy H_t from probability maps $\{\mathbf{p}_i\}$ with Equation 5. Then the reliability map $\mathbf{r}_t$ is estimated by applying a threshold function $\psi_\alpha(\cdot)$ on the Shannon entropy H_t to remain the reliable regions in the final mask prediction for the update of the reliable patch pool.

Experiment

Datasets. We evaluate our model mainly on two widely used VOS benchmarks with multiple objects, YouTube-VOS (Xu et al. 2018a) and DAVIS17 (Pont-Tuset et al. 2017), and a small-scale single object VOS benchmark DAVIS16 (Perazzi et al. 2016). The unseen object categories make

YouTube-VOS more proper to measure the generalization ability of algorithms. So we conduct our experiments on YouTube-VOS to evaluate various methods accurately.

Metrics. We adopt the evaluation metrics from the DAVIS benchmark (Perazzi et al. 2016): the region accuracy J, which calculates the intersection-over-union (IoU) of the estimated masks and the ground truth masks, the boundary accuracy F, which measures the accuracy of boundaries via bipartite matching between the boundary pixels.

Implementation details. We use the DeepLabv3+ (Chen et al. 2018a) architecture as the backbone of our network. Unless otherwise specified, we use ResNet-101 as the backbone network of DeepLabv3+.

The training is conducted with an SGD optimizer with a momentum of 0.9 using the cross-entropy loss. For YouTube-VOS experiments, we only use YouTube-VOS without any external datasets. We first use a learning rate of 0.02 for $200k$ steps with a batch size of 8, then change to a learning rate of 0.01 for another $200k$ steps. During inference, we restrict the long-edge of each frame to no more than 1040 pixels and apply a scale set of [1.0, 1.3, 1.5] for multiscale testing on YouTube-VOS. For DAVIS17 and DAVIS16 experiments, we finetune the trained model on YouTube-VOS for $20k$ steps with both DAVIS and YouTube-VOS in a ratio of 2:1.

All the experiments are performed on an NVIDIA DGX-1 Linux workstation (OS: Ubuntu 16.04.4 LTS, GPU: 8× Tesla V100). Our code is implemented with PyTorch 1.4.1 and is partly leveraged from (Yang, Wei, and Yang 2020). To ensure the validity of experiments, our main results are averaged for 3 runs. For hyper-parameters, we set the update interval τ for reliable patch candidate pool $\mathcal{P}$ as 5 and the parameter α in the reliability filter as 1 without tuning.

Main Results

Quantitative comparison. We compare our method with multiple state-of-the-art methods on YouTube-VOS18 validation (YV18-Val), YouTube-VOS19 validation (YV19-Val) and DAVIS benchmarks in Table 1 and Table 2. With-

Methods	YouTube-VOS 2018 Validation					YouTube-VOS 2019 Validation				
	J&F	J_s	J_u	F_s	F_u	J&F	J_s	J_u	F_s	F_u
AGAME (Johnander et al. 2019)	66.1	67.8	60.8	-	-	-	-	-	-	-
PReM (Luiten, Voigtlaender, and Leibe 2018)	66.9	71.4	56.5	75.9	63.7	-	-	-	-	-
STM$^{\triangle}$ (Oh et al. 2019)	79.4	79.7	72.8	84.2	80.9	-	-	-	-	-
CFBI (Yang, Wei, and Yang 2020)	81.4	81.1	75.3	85.8	83.4	81.0	80.6	75.2	85.1	83.0
RMN (Xie et al. 2021)	81.5	82.1	75.7	85.7	82.4	-	-	-	-	-
SST (Duke et al. 2021)	81.7	81.2	76.0	-	-	81.8	80.9	76.6	-	-
LCM$^{\triangle}$ (Hu et al. 2021)	82.0	82.2	75.7	86.7	83.4	-	-	-	-	-
MiVOS+km$^{\triangle}$ (Cheng, Tai, and Tang 2021b)	82.6	81.1	77.7	85.6	86.2	82.8	81.6	77.7	85.8	85.9
HMMN$^{\triangle}$(Seong et al. 2021)	82.6	82.1	76.8	87.0	84.6	82.5	81.7	77.3	86.1	85.0
CFBI+ (Yang, Wei, and Yang 2021)	82.8	81.8	77.1	86.6	85.6	82.9	80.6	78.9	85.2	86.8
JOINT (Mao et al. 2021)	83.1	81.5	**78.7**	85.9	86.5	82.8	80.8	79.0	84.8	86.6
Ours	**84.0**	**83.1**	78.5	**87.7**	**86.7**	**83.9**	**82.6**	**79.1**	**86.9**	**87.1**
CFBI^{MS,F} (Yang, Wei, and Yang 2020)	82.7	82.2	76.9	86.8	85.0	82.4	81.8	76.9	86.1	84.8
Ours^{MS}	**84.3**	**83.3**	**78.9**	**87.9**	**86.9**	**84.2**	**83.0**	**79.4**	**87.3**	**87.2**

Table 1: Quantitative comparisons on YouTube-VOS. Subscript s and u denote scores in seen and unseen categories. $\triangle$ denotes using external training datasets. Superscript MS and F denotes using multi-scale and flip testing in evaluation respectively.

	CFBI	SST	MiVOS$^{\triangle}$	RMN$^{\triangle}$	LCM	JOINT	**Ours**	**Ours**^{FR}
DAVIS16 Validation (single object, easy)								
J&F	89.4	-	**91.0**	88.8	90.7	-	90.6	**91.5**
J	88.3	-	89.7	88.9	**89.9**	-	87.1	**88.3**
F	90.5	-	92.1	88.7	91.4	-	**94.0**	**94.7**
DAVIS17 Validation (multi-object, medium)								
J&F	81.9	82.5	83.3	83.5	83.5	83.5	**83.7**	**84.8**
J	79.1	79.9	80.6	81.0	80.5	80.8	**81.3**	**82.5**
F	84.6	85.1	85.1	86.0	**86.5**	86.2	86.0	**87.2**
DAVIS17 Test-dev (multi-object, hard)								
J&F	74.8	-	76.5	75.0	78.1	-	**79.2**	**81.0**
J	71.1	-	72.7	71.9	74.4	-	**75.8**	**77.6**
F	78.5	-	80.2	78.1	81.8	-	**82.6**	**84.3**

Table 2: Quantitative comparisons on DAVIS. $\triangle$ denotes using external training datasets besides YouTubeVOS and DAVIS. Superscript FR denotes full-resolution testing. Otherwise, methods are all tested on $480p$.

Method	J&F	J_s	J_u	F_s	F_u
CFBI	81.4	81.1	75.3	85.8	83.4
S2S	81.9(0.5↑)	81.6(0.5↑)	76.1(0.8↑)	86.1(0.3↑)	83.9(0.5↑)
$S2P$	81.9(0.5↑)	81.4(0.3↑)	76.2(0.9↑)	85.9(0.1↑)	84.3(0.9↑)
$P2S$	82.2(0.8↑)	81.8(0.7↑)	76.4(1.1↑)	86.5(0.7↑)	84.3(0.9↑)
$P2P$	82.3(0.9↑)	81.8(0.7↑)	76.3(1.0↑)	86.5(0.7↑)	84.4(1.0↑)
$C2S$	82.1(0.7↑)	81.8(0.7↑)	76.2(0.9↑)	86.3(0.5↑)	84.1(0.7↑)
$S2C$	82.6(1.2↑)	81.7(0.6↑)	**77.5(2.2↑)**	86.1(0.3↑)	**85.2(1.8↑)**
$C2C$	82.3(0.9↑)	81.6(0.5↑)	76.8(1.5↑)	86.1(0.3↑)	84.7(1.3↑)
$C2P$	82.5(1.1↑)	82.0(0.9↑)	76.7(1.4↑)	86.7(0.9↑)	84.4(1.0↑)
$P2C$	**82.9(1.5↑)**	**82.9(1.8↑)**	76.9(1.6↑)	**87.4(1.6↑)**	84.5(1.1↑)
$P\&C$	82.5(1.1↑)	82.1(1.0↑)	76.7(1.4↑)	86.7(0.9↑)	84.3(0.9↑)
$P2P2P$	82.3(0.9↑)	82.0(0.9↑)	76.4(0.9↑)	86.5(0.7↑)	84.1(0.7↑)

Table 3: Ablation on memory modulator scheme variants on YouTube-VOS 18 validation set. ↑ indicates improvement over our compared method CFBI.

out using any bells and whistles (e.g., fine-tuning at test time, top-k filtering, pre-training on external training dataset BL30K or simulated training data), our model significantly outperforms nearly all the contemporary work and previous SOTA methods. Our model stands out in most of the evaluation metrics, especially on unseen categories of YouTube-VOS, which further demonstrates the generalization ability. On the challenging DAVIS17 Test-dev split, the overall performance can be further promoted to 81% J&F with full-resolution testing thanks to the good scalability of input resolution of our model.

Qualitative comparison. Fig. 4 shows the qualitative comparison between state-of-the-art methods and our model on the YouTube-VOS validation set. Thanks to our reliable correction mechanism, our model suppresses the error propagation better, achieving better results when the targets become different from the reference.

Ablation Study and Discussion

In addition to the state-of-the-art performance, we provide the following insights with our detailed ablation.

Modulator assembly variants. Apart from propagation modulator P and correction modulator C, an auxiliary self-modulator S that uses current embedding both as both memory and input serves as a reference. For simplicity, "2" and "&" between characters stand for cascaded and parallel assembly schemes. The variants can be categorized into propagation-based ($S2P$, $P2S$, $P2P$ and $P2P2P$), correction-based ($S2C$, $C2S$ and $C2C$) and propagation & correction ($C2P$, $P2C$ and $P\&C$) schemes. Note that $P\&C$ is a variant where the propagation and correction modulators are assembled in parallel streams and followed by a self-modulator to fuse the outputs.

Modulation with propagation and correction guidance. To study how the propagation and correction guidance influ-
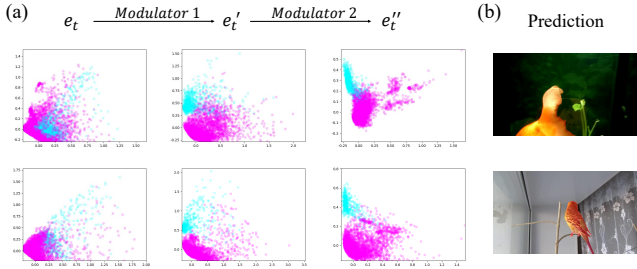
Figure 5: (a) Pixel-level embedding evolution through a $P2C$ modulator. Features of foreground and background are colored in blue and purple. The visualization is conducted by reducing the embedding to 2-dims with PCA. (b) Predicted mask with input frame blended. Zoom in to view better.

$P2C$	OA	RF	J&F	J$_s$	J$_u$	F$_s$	F$_u$
✓			82.7	82.1	77.4	86.3	84.9
✓	✓(W)		82.9	82.3	77.7	86.5	85.2
✓	✓($W+S$)		83.6	82.4	78.7	86.8	86.5
✓	✓($W+S$)	✓	**83.9**	**82.6**	**79.1**	**86.9**	**87.1**

Table 4: Ablation study for reliable object proxy augmentation on YouTube-VOS 19 validation set. $P2C$ denotes the propagation-correction modulator assembly, OA and RF denotes using object proxy augmentation and reliability filter. W denotes using OA for calculate modulation weight only. $W+S$ denotes using OA for calculating weights and correlation (similarity).

ences VOS results, we conduct ablation experiments on the YV18-Val split with different modulator variants. We also list the SOTA model CFBI that uses naive local and global guidance for comparison. The results are shown in Table 3.

Can our modulator improve the performance only with direct propagation guidance? Yes. From Table 3, we can find that our propagation-based variants all outperform CFBI. The reasons are two-fold. First, while CFBI utilizes image-level features from previous frames for improving performance, it does not have explicit designs to suppress error propagation. Instead, our model utilizes a compressed and memorized embedding throughout the whole sequence, which consolidates the local spatial correlation. Secondly, the modulator is a plug-and-play module with a resolution-keeping design, which may help protect the detail from loss.

Can more propagation guidance from different layers further boost the performance? No. Although modulators are light-weight, simply stacking more propagation-based modulators ($P2S$, $P2P$ and $P2P2P$) only has a marginal gain, which indicates that simply incorporating propagation guidance is not enough.

Is correction guidance effective for uncertainty suppression? Yes. Thanks to the highly reliable correction memory, simply leveraging the correction modulator ($C2S$, $S2C$, $C2C$) can boost the performance especially in unseen categories compared with the one without using correction guidance ($S2S$). Among those injections, assembling the correction modulator in deeper layers ($S2C$) brings the largest gain since it encodes cues with high reliability and its correction impact will not be overridden by others.

Can propagation and correction be collaboratively leveraged? Yes. The propagation-correction cascaded assembly ($P2C$) stands out among all the modulator variants, which verifies our insight that propagation modulator relying on local coherence fits shallow layers while correction modulator relying on high-level semantics fits deeper layers.

How does modulator affect the embedding? To answer this question, we visualize the transformation process of embedding in $P2C$ over different layers with principal components analysis in Fig. 5. We can observe that $P2C$ modulates the embedding by progressively separating the foreground and background features in embedding space.

Modulation with reliable object proxy augmentation. To study how the reliable object proxy augmentation influences VOS results, we conduct a set of ablations on YV19-Val split with different proxy construction settings, with $P2C$ as a baseline here. The results are shown in Table 4.

Does object proxy augmentation make contribution? Yes. Table 4 shows that no matter uncertainty patches are filtered or not, using object proxy augmentation always performs better, for both calculating modulation weight and correlation. The reason is that direct propagation may introduce uncertainty while correction from the first frame may be limited due to incomplete semantic guidance. Thus, apart from these two guidance, augmenting object proxy representation especially with reliability helps to complete the semantic concept of an object.

Is reliability-based filtering method beneficial? Yes. Comparing $P2C+OA(W+S)+RF$ and $P2C+OA(W+S)$, we can notice that object proxy augmentation with reliability filter outperforms, which indicates that such consideration of reliability during proxy augmentation can further complement with propagation-correction modulators.

Conclusion

We present a new modulation-based model that can effectively suppress error propagation in semi-supervised VOS. The key is to disentangle the correction from the frame-by-frame mask propagation, which provides reliable and comprehensive object proxies as modulation weights and assembles modulators carefully to further consolidate the target embedding. The object proxy is augmented by supplementing new reliable feature patches from the reliability filter in each iteration, evolving comprehensively. The target embedding encoding the current frame and correlations with references is also consolidated due to the supplemented reliable patches. The assembly of modulators is critical, and our experiments demonstrate that the cascaded propagation-correction scheme performs the best. The main reason is that correction modulation contains global reliable information that could correct errors, and its impact should not be overridden by other modulation.

We also introduce a reliability filter to facilitate the modulation by assessing prediction quality and selecting reliable feature patches. The experiments show impressive gain from the reliable propagation-correction modulation for VOS.

References

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 221–230.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, Y.; Jalali, A.; Sanghavi, S.; and Caramanis, C. 2011. Low-rank Matrix Recovery from Errors and Erasures. arXiv:1104.0354.

Chen, Y.; Pont-Tuset, J.; Montes, A.; and Van Gool, L. 2018b. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1189–1198.

Cheng, H. K.; Tai, Y.-W.; and Tang, C.-K. 2021a. Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion. In *CVPR*.

Cheng, H. K.; Tai, Y.-W.; and Tang, C.-K. 2021b. Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion. *arXiv preprint arXiv:2103.07941*.

Cheng, J.; Tsai, Y.-H.; Hung, W.-C.; Wang, S.; and Yang, M.-H. 2018. Fast and accurate online video object segmentation via tracking parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7415–7424.

Cheng, J.; Tsai, Y.-H.; Wang, S.; and Yang, M.-H. 2017. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, 686–695.

Duke, B.; Ahmed, A.; Wolf, C.; Aarabi, P.; and Taylor, G. W. 2021. SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation. *arXiv preprint arXiv:2101.08833*.

Feder, M.; and Merhav, N. 1994. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1): 259–266.

Goldfeld, Z.; Van Den Berg, E.; Greenewald, K.; Melnyk, I.; Nguyen, N.; Kingsbury, B.; and Polyanskiy, Y. 2019. Estimating Information Flow in Deep Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2299–2308. PMLR.

Hu, L.; Zhang, P.; Zhang, B.; Pan, P.; Xu, Y.; and Jin, R. 2021. Learning Position and Target Consistency for Memory-based Video Object Segmentation. *arXiv preprint arXiv:2104.04329*.

Hu, Y.-T.; Huang, J.-B.; and Schwing, A. 2017. MaskRNN: Instance Level Video Object Segmentation. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hu, Y.-T.; Huang, J.-B.; and Schwing, A. G. 2018. Videomatch: Matching based video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 54–70.

Johnander, J.; Danelljan, M.; Brissman, E.; Khan, F. S.; and Felsberg, M. 2019. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8953–8962.

Khoreva, A.; Benenson, R.; Ilg, E.; Brox, T.; and Schiele, B. 2019. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, 127(9): 1175–1197.

Li, X.; Wang, J.; Li, X.; and Lu, Y. 2021. Hybrid Instance-aware Temporal Fusion for Online Video Instance Segmentation. arXiv:2112.01695.

Li, Y.; Xu, N.; Peng, J.; See, J.; and Lin, W. 2020. Delving into the Cyclic Mechanism in Semi-supervised Video Object Segmentation. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1218–1228. Curran Associates, Inc.

Liang, Y.; Jafari, N.; Luo, X.; Chen, Q.; Cao, Y.; and Li, X. 2020a. WaterNet: An adaptive matching pipeline for segmenting water with volatile appearance. *Computational Visual Media*, 1–14.

Liang, Y.; Li, X.; Jafari, N.; and Chen, J. 2020b. Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refinement. *Advances in Neural Information Processing Systems*, 33.

Liang, Y.; Li, X.; Jafari, N.; and Chen, J. 2020c. Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refinement. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3430–3441. Curran Associates, Inc.

Lin, H.; Qi, X.; and Jia, J. 2019. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3949–3957.

Lu, X.; Wang, W.; Danelljan, M.; Zhou, T.; Shen, J.; and Van Gool, L. 2020. Video object segmentation with episodic graph memory networks. *arXiv preprint arXiv:2007.07020*.

Luiten, J.; Voigtlaender, P.; and Leibe, B. 2018. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, 565–580. Springer.

Mao, Y.; Wang, N.; Zhou, W.; and Li, H. 2021. Joint Inductive and Transductive Learning for Video Object Segmentation. arXiv:2108.03679.

Oh, S. W.; Lee, J.-Y.; Sunkavalli, K.; and Kim, S. J. 2018. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7376–7385.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, 9226–9235.

Perazzi, F.; Khoreva, A.; Benenson, R.; Schiele, B.; and Sorkine-Hornung, A. 2017. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2663–2672.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 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.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Seong, H.; Hyun, J.; and Kim, E. 2020. Kernelized Memory Network for Video Object Segmentation. In *European Conference on Computer Vision*, 629–645. Springer.

Seong, H.; Oh, S. W.; Lee, J.-Y.; Lee, S.; Lee, S.; and Kim, E. 2021. Hierarchical Memory Matching Network for Video Object Segmentation. arXiv:2109.11404.

Tsai, Y.-H.; Yang, M.-H.; and Black, M. J. 2016. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3899–3908.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, u.; and Polosukhin, I. 2017. Attention is All You Need. In *Attention is All You Need*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Voigtlaender, P.; Chai, Y.; Schroff, F.; Adam, H.; Leibe, B.; and Chen, L.-C. 2019. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9481–9490.

Voigtlaender, P.; and Leibe, B. 2017. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*.

Wang, H.; Jiang, X.; Ren, H.; Hu, Y.; and Bai, S. 2021. SwiftNet: Real-time Video Object Segmentation. *arXiv preprint arXiv:2102.04604*.

Wang, W.; Shen, J.; Porikli, F.; and Yang, R. 2018. Semi-supervised video object segmentation with super-trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 41(4): 985–998.

Wang, Z.; Xu, J.; Liu, L.; Zhu, F.; and Shao, L. 2019. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3978–3987.

Wu, R.; Lin, H.; Qi, X.; and Jia, J. 2020. Memory Selection Network for Video Propagation. In *European Conference on Computer Vision*.

Xie, H.; Yao, H.; Zhou, S.; Zhang, S.; and Sun, W. 2021. Efficient Regional Memory Network for Video Object Segmentation. *arXiv preprint arXiv:2103.12934*.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018a. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 585–601.

Xu, Y.-S.; Fu, T.-J.; Yang, H.-K.; and Lee, C.-Y. 2018b. Dynamic video segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6556–6565.

Yang, L.; Wang, Y.; Xiong, X.; Yang, J.; and Katsaggelos, A. K. 2018. Efficient Video Object Segmentation via Network Modulation. *CVPR*.

Yang, Z.; Wei, Y.; and Yang, Y. 2020. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, 332–348. Springer.

Yang, Z.; Wei, Y.; and Yang, Y. 2021. Collaborative Video Object Segmentation by Multi-Scale Foreground-Background Integration. arXiv:2010.06349.

Zeng, X.; Liao, R.; Gu, L.; Xiong, Y.; Fidler, S.; and Urtasun, R. 2019. Dmm-net: Differentiable mask-matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3929–3938.