FlowVOS: Weakly-Supervised Visual Warping for Detail-Preserving and Temporally Consistent Single-Shot Video Object Segmentation

Julia Gong
jxgong@cs.stanford.edu
F. Christopher Holsinger
holsinger@stanford.edu
Serena Yeung
syyeung@stanford.edu

Stanford University
Stanford, California, USA

Abstract

We consider the task of semi-supervised video object segmentation (VOS). Our approach mitigates shortcomings in previous VOS work by addressing detail preservation and temporal consistency using visual warping. In contrast to prior work that uses full optical flow, we introduce a new foreground-targeted visual warping approach that learns flow fields from VOS data. We train a flow module to capture detailed motion between frames using two weakly-supervised losses. Our object-focused approach of warping previous foreground object masks to their positions in the target frame enables detailed mask refinement with fast runtimes without using extra flow supervision. It can also be integrated directly into state-of-the-art segmentation networks. On the DAVIS17 and YouTubeVOS benchmarks, we outperform state-of-the-art offline methods that do not use extra data, as well as many online methods that use extra data. Qualitatively, we also show our approach produces segmentations with high detail and temporal consistency.

1 Introduction

Video object segmentation (VOS) has become an increasingly studied task in the computer vision community. The goal of VOS is to label each pixel of each frame of a video with a corresponding class—either one of potentially several foreground objects, or the background. In particular, the semi-supervised inference setting of this task provides the ground-truth segmentation mask for the first video frame, and methods aim to segment these objects for all subsequent frames. This task is difficult because objects in motion can move and deform in different ways, not to mention additional challenges such as camera motion and occlusions.

Many deep learning-based methods have been proposed to tackle this problem. Recent state-of-the-art offline methods [12, 13] have used a memory bank of encoded previous frames and masks, which is queried when segmenting later frames. The advantage of these approaches is that the learned latent spaces robustly encode higher-level features of the target objects from previous frames; however, they lack the detail to propagate fine features and movements across consecutive frames and thus can suffer from temporal inconsistency.

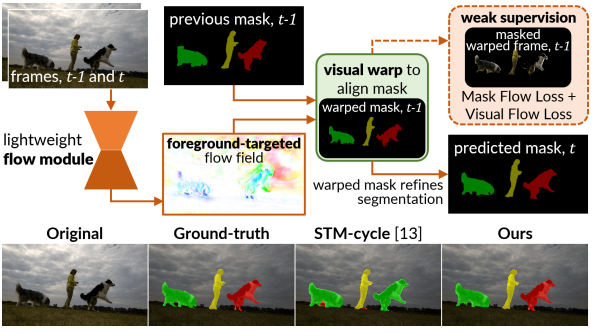


Figure 1: We introduce a weakly-supervised visual warping approach for VOS that improves detail and temporal consistency. Our lightweight flow module learns to regress a foreground-targeted flow field that warps the previous mask from time $t-1$ to align it to frame t. To learn detailed flow fields, we directly exploit the VOS data using two weakly-supervised losses, rather than learning the cumbersome general optical flow task. We show that our warped masks effectively refine the final segmentations. Our method can be easily integrated into state-of-the-art segmentation models, does not require extra data, and has fast frame rates.

Our work's key insight is to mitigate these issues of detail preservation and temporal consistency using *visual warping*, which captures small deviations between video frames. As shown in Figure 1, our method warps the object masks from previous frames toward the target frame to add detail and temporal consistency to the final segmentation. To learn these deltas between frames, we introduce a weakly-supervised flow module that can be easily used with state-of-the-art segmentation networks. Some early VOS works employed traditional optical flow methods to perform visual warping [20, 23], but they use highly time-costly online optical flow optimization methods. More recent VOS works [9, 14, 15] incorporate warping by using state-of-the-art deep learning optical flow estimation networks like FlowNet 2.0 [10]. However, these networks have high computational cost and require extensive pretraining on supervised data focused on the full optical flow task [6, 17]. Their ability to predict detailed motions also degrades significantly even with small speed increases.

In this work, we address these shortcomings by proposing a novel visual warping approach for VOS. Unlike prior works that approach visual warping using standalone, pretrained optical flow methods, we do not predict traditional optical flow. Instead, we directly train on the VOS data of interest to learn an offline flow module for VOS-specific visual warping. Our foreground-targeted approach focuses on aligning previous foreground objects to their new positions. Specifically, we introduce two weakly-supervised flow losses that enforce pixel-level consistency between warped previous masks and frames and the target masks and frames. These train the flow module to capture small changes between timesteps in pixel-wise flow fields, which warp previous object masks to propagate detail to the final prediction. Moreover, since we do not learn the general optical flow task, our flow module can stay lightweight for fast frame rates, and by training directly on the VOS data, we do not need any supervised flow data. As such, our method can generalize to diverse data not studied by traditional optical flow techniques, and more generally to object-focused scenes.

On the two major VOS benchmarks, DAVIS17 [20] and YouTubeVOS [29], our method achieves state-of-the-art performance among offline works that do not use extra training data (e.g. additional datasets, image segmentations, supervised optical flow, or synthetic data). Additionally, we outperform or stay competitive with those that use online learning and

extra data. We also qualitatively show that our method achieves greater segmentation detail preservation and temporal consistency. Our contribution can be summarized as follows:

- We propose a novel foreground-targeted visual warping approach that improves segmentation detail and temporal consistency for VOS. We show that instead of learning traditional optical flow, our flow module jointly learns detail-preserving flow fields by exploiting the target VOS data directly using two weakly-supervised losses.
- Our purely offline-learned flow module for VOS is fast and can be easily integrated into state-of-the-art segmentation networks (here, we integrate it into STM-cycle [13]).
- On DAVIS17 [21] and YouTubeVOS [29], we achieve state-of-the-art performance among works that do not use online learning nor extra data. We also outperform or stay competitive with those that do, while maintaining faster frame rates.

2 Related Work

Semi-Supervised Video Object Segmentation. With the success of deep learning, semi-supervised VOS has seen a large number of works in recent years. These works leverage a variety of strategies, including online versus offline optimization, mask propagation, segmentation by tracking, coarse-to-fine refinement, and usage of attention and memory banks.

Semi-supervised VOS methods can be considered *online* or *offline*, where inference includes learning in the former and does not in the latter. Online techniques are often used for mask refinement. OSVOS [2] introduced the first deep learning-based online VOS method, which gradually refines the model from segmenting general objects to those in the initial reference mask; OnAVOS [25] adds an adaptive learning mechanism. MaskTrack [20]'s online learning method learns mask refinement from external static images. PReMVOS [15] achieves strong performance via coarse-to-fine refinement of object proposals and optical flow, though it is among the most computationally intensive. More recent work also employs online learning on the initial reference mask to refine the prediction [9]. While online learning methods can produce detailed segmentations, their high computational cost causes slow inference frame rates impractical for real-time settings.

In contrast, offline learning methods do not update during inference, generally yielding faster frame rates. Our method lies in this category. Recently, D3S [16] proposed segmenting frames independently with explicit foreground-background separation, though its temporal consistency drops in multi-object settings. To enforce consistency, some offline methods leverage previous frames; S2S [29] and RVOS [24] use recurrent networks, while RGMP [19] use Siamese encoders for the previous and current frame. The weakness of these methods is that they do not target specific features across frames. Many methods [12, 13, 14, 26, 28] therefore use attention mechanisms to achieve stronger performance.

One such work, FEELVOS [26], leverages the initial reference mask for explicit feature matching with subsequent frames. AGSS-VOS [14] further attends over the previous frame and mask, while STCNN [28] attends over multi-scale feature maps. STM [12] achieves even higher performance by introducing an external memory bank and attending over multiple previous frames by querying them using the target frame. State-of-the-art STM-cycle [13] adds a cyclic loss to reduce error propagation. Concurrently, some works [7, 18] instead use attention via transformers to address pixel spatiotemporal relations and model scalability. While attention works capture higher-level features of foreground objects, their segmentations often lack detail or fail to propagate fine deformations and movements across frames, even when the object does not change much. Our insight is that explicitly learning the differ-

ences between pairs of frames can address these issues. Thus, improving on previous work, we leverage visual warping to add temporal consistency and segmentation detail.

Optical Flow Estimation in VOS. Previous VOS works that perform visual warping all use optical flow prediction methods to do so. Early VOS works such as [23] and MaskTrack [20] use traditional online optical flow estimation methods. Later work uses deep learning-based optical flow prediction approaches, such as FlowNet [6], which pioneered the task as a supervised problem using convolutional neural networks. Using [6], SegFlow [3] jointly learns optical flow and segmentation with deep learning, but requires online learning and learns the full optical flow task, thus requiring supervised flow annotations. Works such as [3, 20, 23] are either online, use online optical flow estimation, or both, rendering them computationally costly and unfit for real-time settings. [3, 20] also require extra data for training, and they do not exploit the warped masks themselves to guide the segmentation.

More recent VOS work uses state-of-the-art optical flow prediction networks that outperform FlowNet [6]; notably, FlowNet 2.0 [10] achieves a significant improvement by stacking multiple convolutional networks end-to-end. While [10] is state-of-the-art, it requires extensive pretraining on several supervised optical flow datasets [6, 17] and is computationally costly to stack on top of a segmentation model. Moreover, without extensive ensembling of multiple networks end-to-end, its ability to predict detailed movements drops. Still, many strong VOS works accept these tradeoffs and use [10]. In particular, recent works PReMVOS [15] and AGSS-VOS [14] use the pretrained FlowNet 2.0 [10] to predict optical flow, which they use to warp previous masks either for further refinement or to guide attention mechanisms. However, these works suffer from FlowNet 2.0 [10]'s accuracy-speed tradeoff; they use the FlowNet 2.0 ensemble at the cost of significant frame rate drops. Moreover, they rely on the extensive extra flow data used to pretrain FlowNet 2.0, which generates additional complexity and data requirements, while importantly not being foreground object-targeted.

Our approach mitigates these issues in prior work. Specifically, our key insight is to learn a weakly-supervised, foreground-targeted visual warping model for VOS instead of learning general optical flow. Even with no extra data and at faster speeds, our approach produces detailed and temporally consistent segmentations and can train directly on the target data.

3 Methods

Our method tackles semi-supervised video object segmentation (VOS), improving the detail and temporal consistency of prior work. The key contribution of our approach is the weakly-supervised flow module, which learns to regress foreground object-targeted flow fields that warp previous masks toward objects in the target frame to preserve fine detail and temporal consistency. As it need not learn general optical flow, it can remain lightweight. It can be used with any segmentation network to refine masks; we use STM-cycle [13] in this work. To train the flow module, we introduce two weakly-supervised losses for flow field regression, which encourage pixel-level consistency between warped previous object masks and frames and the target masks and frames. Unlike prior work [14, 15], these losses exploit the VOS data to regress object-targeted flows without any extra supervision from ground-truth flow fields. Below, we review the semi-supervised VOS problem definition, describe our model architecture, formalize the two weakly-supervised flow losses, and discuss model training.

We first define notation for semi-supervised VOS. Given a video with T frames, let the t-th frame in temporal order be X_t and its corresponding ground-truth mask be Y_t, for $t \in [1,T]$. In training, all ground-truth masks are provided; in inference, only the first ground-truth mask Y_1 is provided. The goal of the model is to predict masks $\hat{Y}_t$ for all subsequent frames.

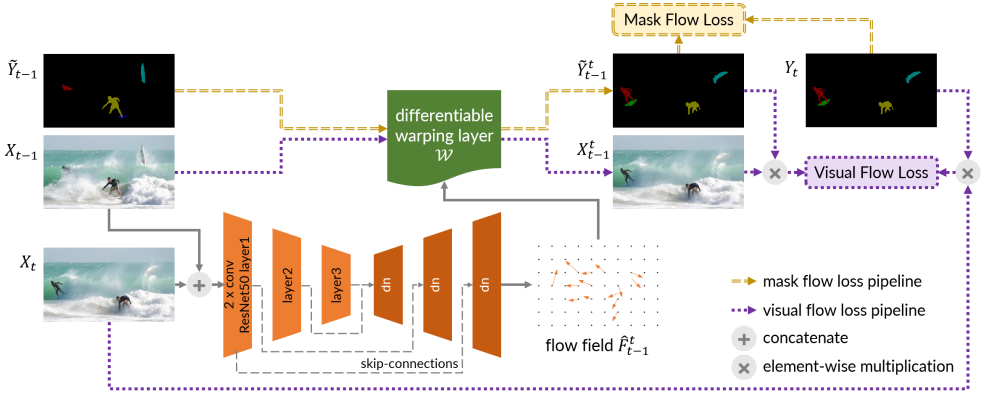


Figure 2: Overview of the proposed flow module. Given a previous and target frame (here shown as X_{t-1}, X_t), it regresses a flow field to warp the previous frame and mask to the current ones. The previous mask $\tilde{Y}_{t-1}$ can be either the ground-truth or predicted mask from timestep $t-1$ in training (Sec. 3.4), but only the predicted mask in inference. Two weakly-supervised losses train the flow module: the Mask Flow Loss minimizes the difference between the warped previous mask $\tilde{Y}^t_{t-1}$ and target mask Y_t (yellow double-dashes), and the Visual Flow Loss minimizes the difference between the warped previous frame X^t_{t-1} masked by $\tilde{Y}^t_{t-1}$ and the target frame X_t masked by Y_t to eliminate background noise (purple dotted lines).

3.1 Flow Module

As shown in Figure 2, we introduce a lightweight flow module $\mathcal{F}$ to address detail preservation and temporal consistency. $\mathcal{F}$ is an hourglass network to enable learning features of different scales (Implementation Details in Sec. 4.2). Given a pair of frames from the same video, $\mathcal{F}$ generates a flow field that describes object movement between the previous and current frame. For a target frame X_t, the flow module takes two frames in temporal sequential order, here $\{X_{t-1}, X_t\}$, and an object mask $\tilde{Y}_{t-1}$ for the previous frame (in training either the ground-truth Y_{t-1} or predicted mask $\hat{Y}_{t-1}$ (Sec. 3.4); in inference only the latter). It outputs a flow field $\hat{F}^t_{t-1}$ that warps X_{t-1} to X_t and $\tilde{Y}_{t-1}$ to Y_t. The flow field has the same spatial dimensions as the video frames, with two channels corresponding to pixel-wise x and y displacements normalized by frame width and height. The flow module's function is thus

$$\hat{F}^t_{t-1} = \mathcal{F}(X_{t-1}, X_t, \tilde{Y}_{t-1}). \tag{1}$$

This flow field $\hat{F}^t_{t-1}$ warps the previous object mask $\tilde{Y}_{t-1}$ toward the target mask using the differentiable warping layer $\mathcal{W}$ introduced in Spatial Transformer Networks [□], which transforms images using a sampling kernel. The resulting warped previous mask $\tilde{Y}^t_{t-1}$ is thus

$$\tilde{Y}^t_{t-1} = \mathcal{W}(\tilde{Y}_{t-1}, \hat{F}^t_{t-1}). \tag{2}$$

We train the flow module to regress effective visual warping flow fields with only VOS data and no extra flow data. To do so, we introduce two weakly-supervised losses.

3.2 Weakly-Supervised Flow Losses

Our work contributes a novel weakly-supervised approach to regress visual warping flow fields for VOS. In contrast to prior works that use heavy pretrained optical flow models,

our key insight is to leverage existing VOS data to directly learn flow fields in a weakly-supervised manner using two losses: the Mask Flow Loss (MFL) and Visual Flow Loss (VFL). Both train the flow module to warp previous objects toward the targets by penalizing pixel-level differences after warping. The MFL minimizes differences between warped and ground-truth target masks, while the VFL does the same for warped and target video frames.

Mask Flow Loss (MFL). Our goal is to align previous object masks to the target frame. As such, we require the warped previous mask $\tilde{Y}^t_{t-1}$ (Eq. 2) to be close to the target mask Y_t. The MFL minimizes the difference between the warped previous mask and target mask. It combines the commonly-used cross-entropy and mask IOU losses between these two masks,

$$\mathcal{L}_{MF} = \frac{1}{|P|}\sum_{u\in P}\left((1-Y_{t,u})\log(1-\tilde{Y}^t_{t-1,u}) + Y_{t,u}\log(\tilde{Y}^t_{t-1,u})\right) - \lambda\frac{\sum_{u\in P}\min(\tilde{Y}^t_{t-1,u}, Y_{t,u})}{\sum_{u\in P}\max(\tilde{Y}^t_{t-1,u}, Y_{t,u})}, \tag{3}$$

where P is the set of pixel coordinates, $Y_{t,u}$ and $\tilde{Y}^t_{t-1,u}$ are the mask pixel values at coordinate u for the ground-truth and warped previous masks respectively, and λ weights the two losses. This combines their strengths: cross-entropy favors pixel-level accuracy and optimizes more stably, while the IOU loss enforces overall object shape and better handles class imbalances.

Visual Flow Loss (VFL). To warp previous objects to the target frame, we can similarly exploit visual appearance; we thus also require the warped previous frame denoted X^t_{t-1} (achieved via the analogous operation on X_{t-1} using Eq. 2) to be close to the target frame X_t. A naive formulation of the VFL may be the pixel-wise mean squared error (MSE) between the warped previous frame and the target frame. However, the disadvantage of this formulation is that video frames can exhibit large amounts of visual noise in the background that is irrelevant to the objects of interest. This could be caused by background activity, camera motion, occlusions, or motion blur, among other factors.

Thus, since we are only concerned with the motion of the foreground objects for the VOS task, a stronger Visual Flow Loss will only take these pixels into consideration to more precisely capture object movement. Specifically, we use the MSE loss between two *masked* frames: the target frame X_t masked with the target mask Y_t, and the warped previous frame X^t_{t-1} masked with the warped previous mask $\tilde{Y}^t_{t-1}$ to obtain (continuing the notation in Eq. 3):

$$\mathcal{L}_{VF} = \frac{1}{|P|}\sum_{u\in P}\left((X^t_{t-1,u})(\tilde{Y}^t_{t-1,u}) - (X_{t,u})(Y_{t,u})\right)^2. \tag{4}$$

3.3 End-to-End Segmentation Method

Our flow module can be easily integrated into state-of-the-art segmentation networks; here, we present a version of our method that uses STM-cycle [13], which is based on STM [12]. Our end-to-end method, shown in Figure 3, uses the flow module's output flow field to refine the final segmentation. To integrate our flow module into the segmentation network, the regressed flow field first warps the previous mask $\tilde{Y}_{t-1}$ to yield $\tilde{Y}^t_{t-1}$, as discussed in Eq. 2. We concatenate this warped previous mask with the output of the second-to-last convolutional block of the decoder $\mathcal{D}$. The last convolutional block of $\mathcal{D}$ outputs the final predicted segmentation $\hat{Y}_t$. (Note that networks without decoders can use a structure with a similar purpose, such as a refinement module). The segmentation network loss is a combination of the mask IOU and cross-entropy losses (as in Eq. 3), encouraging the final segmentation to be close to the ground-truth mask Y_t. See Supplementary for further details.

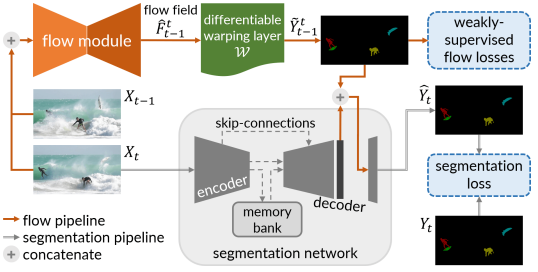


Figure 3: End-to-end FlowVOS method. Our flow module takes the previous and target frames X_{t-1}, X_t as input. We train it to generate a flow field $\hat{F}^t_{t-1}$ that warps a previous mask to align it with the target frame X_t using two weakly-supervised losses. The warped mask $\tilde{Y}^t_{t-1}$ is concatenated to the segmentation decoder's second-to-last convolutional block feature map to predict the final mask $\hat{Y}_t$. In training, the weakly-supervised losses are used together with a standard segmentation loss (mask IOU and cross-entropy, as in Eq. 3).

3.4 Training

As the flow module learns only from weakly-supervised signals, it produces distorted flow fields when given unreasonable previous masks. This can occur early in training or during simple joint training of the flow module and segmentation network, when predicted masks are noisy. To overcome these challenges, we use two training mechanisms that stabilize learning: previous mask teacher-forcing and two-stage training. We analyze hyperparameter robustness and comment further on these two mechanisms in the Supplementary.

Previous mask teacher-forcing. Teacher-forcing [1] is a sampling technique widely used in autoregressive model training, where a model's own output is used for its next prediction during inference. As noted in Sec. 3.1, our flow module warps the previous predicted mask in inference. Thus, we use teacher-forcing in training to allow it to learn good flow fields from warped ground-truth masks as well. In training only, the previous mask $\tilde{Y}_{t-1}$ warped by the flow module is teacher-forced with the ground-truth mask Y_{t-1} with probability p, and the network's previous predicted mask $\hat{Y}_{t-1}$ with probability $1-p$.

Two-stage training. We alternately freeze the segmentation model weights to let both modules learn from the progress of the other. [3] used a similar strategy when training their joint models. First, we train the segmentation model $\mathcal{S}$ for E_s epochs. We then add the flow module $\mathcal{F}$, which stays unfrozen for the rest of training. We alternately unfreeze and freeze $\mathcal{S}$ every E_a epochs until convergence.

4 Experiments

4.1 Datasets and Evaluation Metrics

We train and evaluate on two major benchmark datasets for the VOS task: DAVIS17 [21] and YouTubeVOS [29]. For reproducibility, our code and models will be made available online.

DAVIS 17. DAVIS17 [21] has 120 videos total, with a maximum of 10 objects per video. Using the official dataset splits, we train on the 60 training videos and evaluate on the 30 validation and 30 test-dev videos. We use the official DAVIS17 evaluation protocol [21], which is the Jaccard (IOU) mean $\mathcal{J}$ and contour F-score $\mathcal{F}$ across all objects and videos, as well as the mean of $\mathcal{J}$ and $\mathcal{F}$ for the overall score.

OL & ED methods	ED	OL	$\mathcal{J}\%$	$\mathcal{F}\%$	$\mathcal{J}\&\mathcal{F}\%$	FPS
STCNN [18]	✓	✓	58.7	64.6	61.7	0.26†
OnAVOS [36]	✓	✓	64.5	71.2	67.9	0.1
BoLTVOS [13]	✓	✓	72.0	80.6	76.3	0.69
TANDTM [1]	✓	✓	72.3	79.4	75.9	7.1
PReMVOS^F [24]	✓	✓	73.9	81.7	77.8	0.03
OSMN [34]	✓	-	52.5	57.1	54.8	8
RGMP [20]	✓	-	64.8	68.6	66.7	3.6
AGSS-VOS^F [19]	✓	-	64.9	69.9	67.4	10
DMM-Net [25]	✓	-	68.1	73.3	70.7	-
FEELVOS [30]	✓	-	69.1	74.0	71.5	2
STM [23]	✓	-	79.2	84.3	81.8	6.3
OSVOS [3]	-	✓	64.7	71.3	68.0	0.1
STM-cycle [13]	-	✓	69.3	75.3	72.3	9.3
non-OL, non-ED	ED	OL	$\mathcal{J}\%$	$\mathcal{F}\%$	$\mathcal{J}\&\mathcal{F}\%$	FPS
FAVOS [6]	-	-	54.6	61.8	58.2	0.8
D3S [21]	-	-	57.8	63.8	60.8	25
STM-cycle [13]	-	-	68.7	74.7	71.7	**31.9***
Ours	-	-	**70.6**	**75.8**	**73.2**	17.3

(a)

OL & ED methods	ED	OL	$\mathcal{J}\%$	$\mathcal{F}\%$	$\mathcal{J}\&\mathcal{F}\%$	FPS
OnAVOS [36]	✓	✓	53.4	59.6	56.9	0.03
TANDTM [1]	✓	✓	61.3	70.3	65.4	7.1
PReMVOS^F [24]	✓	✓	67.5	75.7	71.6	0.02
RGMP [20]	✓	-	51.3	54.4	52.8	2.4
AGSS-VOS^F [19]	✓	-	54.8	59.7	57.2	9
FEELVOS [30]	✓	-	55.2	60.5	57.8	1.8
STM-cycle [13]	-	✓	55.3	62.0	58.6	6.9
non-OL, non-ED	ED	OL	$\mathcal{J}\%$	$\mathcal{F}\%$	$\mathcal{J}\&\mathcal{F}\%$	FPS
RVOS [29]	-	-	48.0	52.6	50.3	22.7
STM-cycle [13]	-	-	55.1	60.5	57.8	**25.9***
Ours	-	-	**57.1**	**63.1**	**60.1**	13.7

(b)

OL & ED methods	ED	OL	$\mathcal{G}\%$	$\mathcal{J}_\mathcal{S}\%$	$\mathcal{J}_\mathcal{U}\%$	$\mathcal{F}_\mathcal{S}\%$	$\mathcal{F}_\mathcal{U}\%$	FPS
MaskTrack [18]	✓	✓	53.1	59.9	45.0	59.5	47.9	0.05
OnAVOS [36]	✓	✓	55.2	60.1	46.6	62.7	51.4	0.05
DMM-Net [25]	✓	✓	58.0	60.3	50.6	63.5	57.4	-
PReMVOS^F [24]	✓	✓	66.9	71.4	56.5	75.9	63.7	0.17
BoLTVOS [13]	✓	✓	71.1	71.6	64.3	-	-	0.74
OSMN [34]	✓	-	51.2	60.0	40.6	60.1	44.0	4.2
DMM-Net [25]	✓	-	51.7	58.3	41.6	60.7	46.3	12
RGMP [20]	✓	-	53.8	59.5	-	45.2	-	7
AGSS-VOS^F [19]	✓	-	71.3	71.3	65.5	75.2	73.1	12.5
STM [23]	✓	-	79.4	79.7	72.8	84.2	80.9	6.3
OSVOS [3]	-	✓	58.8	59.8	54.2	60.5	60.7	0.06
S2S [33]	-	✓	64.4	71.0	55.5	70.0	61.2	0.06
STM-cycle [13]	-	✓	70.8	72.2	62.8	76.3	71.9	13.8
non-OL, non-ED	ED	OL	$\mathcal{G}\%$	$\mathcal{J}_\mathcal{S}\%$	$\mathcal{J}_\mathcal{U}\%$	$\mathcal{F}_\mathcal{S}\%$	$\mathcal{F}_\mathcal{U}\%$	FPS
RVOS [29]	-	-	56.8	63.6	45.5	67.2	51.0	24
S2S [33]	-	-	57.6	66.7	48.2	65.5	50.3	6
STM-cycle [13]	-	-	69.9	**71.7**	61.4	**75.8**	70.4	**30.3***
Ours	-	-	**71.1**	**71.7**	**64.0**	75.2	**73.3**	16.7

(c)

Table 1: Comparison with state-of-the-art methods on DAVIS17 validation (a), DAVIS17 test-dev (b), and YouTubeVOS validation (c). 'ED' denotes usage of extra training data. 'OL' denotes online learning. Superscript 'F' denotes usage of optical flow. In (a), † denotes runtimes only available on DAVIS16. In (c), $\mathcal{S}, \mathcal{U}$ subscripts denote classes seen and unseen in training, and $\mathcal{G}$ is the global mean. Other method results in (c) taken from [12, 13].

YouTubeVOS. YouTubeVOS [29] is the largest VOS dataset to-date, with a maximum of 12 objects per video. We use the official dataset splits, with 3,471 training videos (training) and 474 validation videos (evaluation). The evaluation protocol also averages $\mathcal{J}$ and $\mathcal{F}$ scores over seen and unseen classes separately, which are then averaged for the overall score.

4.2 Implementation Details

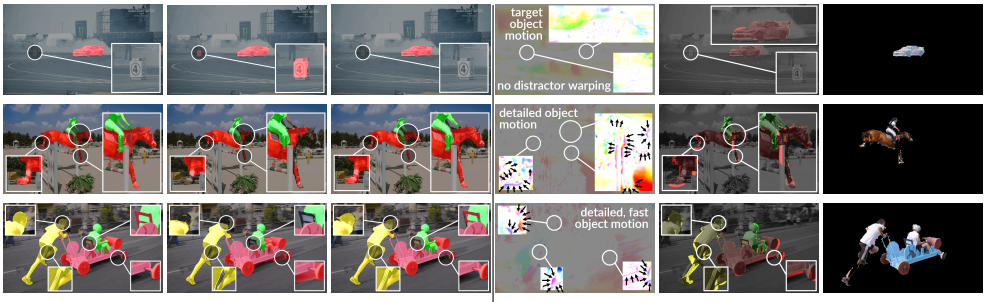
Model. Our flow module's lightweight U-Net [22] structure uses a ResNet50 [8] encoder pretrained on ImageNet [5] and skip-connections between symmetric encoder-decoder blocks. Decoder layers use bilinear upsampling, followed by two 3×3 convolutions. We integrate our module into the recent state-of-the-art STM-cycle [13].* Following training procedures in [12, 13, 26], we pool DAVIS17 and YouTubeVOS training splits in all experiments. See Supplementary for training details (e.g. hyperparameters, hardware, and data augmentation).

4.3 Quantitative Results

We compare our method's performance and speed against state-of-the-art works on DAVIS17 [21] and YouTubeVOS [29]. These include works that use extra annotated data (ED), online learning (OL), and optical flow (shown as superscript F). Tables 1(a), 1(b), and 1(c) respectively show DAVIS17 validation, DAVIS17 test-dev, and YouTubeVOS validation results.

DAVIS17. As shown in Tables 1(a) and 1(b), on the official validation and test-dev splits, we achieve state-of-the-art performance among methods that do not use extra data (ED) nor online learning (OL) (and many that do), while maintaining high frame rates. We achieve +1.9% and +2.0% $\mathcal{J}$ gains over [13] respectively. Notably, on both splits, we outperform STM-cycle [13]'s online learning version, even though it optimizes for performance

*We ran STM-cycle [13] using the authors' provided code and could not replicate the reported frame rates in [13]. With our TITAN RTX GPU, which is faster than the TITAN Xp in [13], we still only achieved the FPS in the tables. For fair comparison, since [13] is the closest work to ours, we report our replicated FPS on the same GPU we used to benchmark our model. [13]'s reported FPS corresponding to Tables 1(a), 1(b), and 1(c) are 38, 31, and 43, respectively. Note that the reported YouTubeVOS validation set result in Table 2 of [13] also incorrectly switched the STM [12] $\mathcal{F}_\mathcal{S}, \mathcal{J}_\mathcal{U}$ scores. We show the correct version from [12] here in Table 1(c).



(a) Ground-truth (b) STM-cycle [13] (c) Ours (c1) Flow field (c2) Warp diff. (c3) Warped frame

Figure 4: Qualitative comparison with STM-cycle [13] on DAVIS17 validation (a, b, c), and our method's intermediate outputs (c1-3). Following [10], we color-code flow fields (c1) with polar coordinate displacements. (c2) brightens pixels that exist in the previous, but not the warped mask, highlighting motion that corresponds to the flows. The masked warped frames (c3) show that our warping operation accurately preserves object detail. In row 1, our flow (c1) captures the target car's leftward (green, blue) motion, but not the distractor due to the foreground-focused losses. In row 2, our detailed flow field propagates object details, such as warping the back hoof upward (purple). In row 3, even with fast-moving objects, the flow accurately warps boundaries of details like the lower right wheel and upper left head.

by adding time-costly iterative mask refinement. This demonstrates that our flow module's warped masks can replace detailed mask refinement without sacrificing speed. Among works that use optical flow, on validation, we achieve a +5.8% $\mathcal{J}\&\mathcal{F}$ improvement over AGSS-VOS [14] and stay competitive with PReMVOS [15] with significantly faster speeds.

YouTubeVOS. YouTubeVOS is the largest VOS benchmark. Since it evaluates performance on classes unseen in training, it measures generalization well. As shown in Table 1(c), on the official validation set, we achieve state-of-the-art performance among methods that do not use extra data (ED) nor online learning (OL), while maintaining a high frame rate. We also outperform all but two works that use ED, OL, or both, which both have lower frame rates. Crucially, our method generalizes significantly better to unseen classes than both offline (+2.6% $\mathcal{J}_\mathcal{U}$, +2.9% $\mathcal{F}_\mathcal{U}$) and online versions of STM-cycle [13], showing that our foreground-targeted approach learns motion priors to better segment unseen objects. We outperform the optical flow-equipped PReMVOS [15] by +7.5% $\mathcal{J}_\mathcal{U}$ with 98 times the speed, highlighting the strengths of our visual warping compared to traditional optical flow.

4.4 Ablation Analysis

In the top half of Table 2, we analyze relative contributions of key components of our method. Our score drops 1.0% with either just the Mask Flow Loss or Visual Flow Loss (VFL); this shows the benefit of leveraging both warped masks and frames in our method. Without masking the warped frame in the VFL, our score is similar to not using the VFL at all, showing the importance of masking the foreground to eliminate background noise, such as the distractors in the first row of Figure 4. In the bottom half of Table 2, we show the importance of using the warped previous mask for mask refinement. When we replace the input to the decoder with the previous predicted mask or video frame, the network performs on-par with STM-cycle [13], meaning these inputs provide less useful information for refining the mask. With the flow field, performance still lacks by 0.6%, showing the benefit of explicitly warping the

Ablations		$\mathcal{J}\&\mathcal{F}\%$
STM [13] (same train protocol as [13], no extra data)		70.5
STM-cycle [13]	+Cycle consistency loss	71.7
Weakly-supervised flow module losses	+Mask Flow Loss only	72.2
	+Visual Flow Loss only	72.2
Foreground masking	VFL w/o foreground masking	72.4
Alternative inputs to segmentation decoder		
Previous predicted mask		71.6
Previous video frame		71.9
Flow field		72.6
Previous masked warped frame		73.0
Ours		**73.2**

Table 2: Ablation study of our method components on DAVIS17 validation. On top, we show ablations using only the Mask Flow Loss (MFL) or Visual Flow Loss (VFL), and without foreground masking of the VFL. The bottom shows alternative segmentation decoder inputs instead of the warped previous mask that our method uses.

previous mask. The previous masked warped frame expectedly performs on-par with our method, since we warp it identically to the previous mask; still, this shows that the warped mask provides stronger signal about object motion.

4.5 Qualitative Results

In contrast to prior works, we show that our foreground-targeted approach for VOS-specific visual warping produces detailed flow fields without needing to learn the general optical flow task. Figure 4 illustrates our method's improvements over the state-of-the-art STM-cycle [13] in segmentation detail (see Supplementary figures for temporal consistency).

In Figure 4, we show that our method preserves segmentation detail in cases with background distractors (row 1), small object details (2), and fast-moving objects (3). Notice that our flow fields (c1) are detailed and correspond to the pixel-wise differences between the previous and warped mask highlighted in (c2). Note that since we do not learn general optical flow (which has ground-truth), there is more than one possible flow field that can accurately warp a foreground object in our VOS-specific setting. This means we can learn detailed motion with greater flexibility; for instance, our model often warps background pixels to achieve better foreground alignment (e.g. the car boundary in row 1), which traditional optical flow would penalize, but our weakly-supervised losses do not. Our visual warping method also enables stronger temporal consistency despite diverse challenges (see Supplementary).

5 Conclusion

We propose a novel foreground-targeted visual warping approach that improves segmentation detail and temporal consistency for semi-supervised video object segmentation (VOS). Instead of learning full optical flow, our flow module learns detailed flow fields using two weakly-supervised losses that directly leverage the target VOS data, which could benefit diverse use cases. We show that the resulting warped masks from our method effectively refine the final segmentations. Our module can be easily integrated into state-of-the-art segmentation networks. Since it does not predict full optical flow, it is lightweight, fast, and requires no extra training data. On the DAVIS17 and YouTubeVOS benchmarks, we achieve state-of-the-art performance among methods that do not use online learning nor extra data. We also outperform or stay competitive with those that do, while maintaining faster frame rates.

Acknowledgements

This work was partially supported by a grant from the Isackson Family Fund for Research in Head and Neck Surgery for compute and support for authors J.G. and F.C.H. The authors also thank Joy Hsu, Shih-Cheng Huang, Rui Yan, Jeffrey Gu, Ali Mottaghi, Nikita Bedi, Danfei Xu, Geeticka Chauhan, and Benjamin Newman for their thoughts, suggestions, and support.

References

[1] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, December 2015.

[2] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixe, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[7] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W. Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5912–5921, June 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Xuhua Huang, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Fast video object segmentation with temporal aggregation network and dynamic template matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

[12] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] Yuxi Li, Ning Xu, Jinlong Peng, John See, and Weiyao Lin. Delving into the cyclic mechanism in semi-supervised video object segmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1218–1228. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0d5bd023a3ee11c7abca5b42a93c4866-Paper.pdf.

[14] Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[15] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, 2018.

[16] Alan Lukezic, Jiri Matas, and Matej Kristan. D3s - a discriminative single shot segmentation tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Jianbiao Mei, Mengmeng Wang, Yeneng Lin, and Yong Liu. Transvos: Video object segmentation with transformers. *arXiv:2106.00588*, 2021.

[19] 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.

[20] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung. Learning video object segmentation from static images. In *Computer Vision and Pattern Recognition*, 2017.

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

[22] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[23] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[25] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *British Machine Vision Conference*, 2017.

[26] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Paul Voigtlaender, Jonathon Luiten, and Bastian Leibe. Boltvos: Box-level tracking for video object segmentation. *arXiv:1904.04552*, 2019.

[28] Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, and Qingming Huang. Spatiotemporal cnn for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[30] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[31] X. Zeng, R. Liao, L. Gu, Y. Xiong, S. Fidler, and R. Urtasun. Dmm-net: Differentiable mask-matching network for video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.