Memory Selection Network for Video Propagation

Ruizheng Wu[1⋆], Huaijia Lin[1⋆], Xiaojuan Qi[2], and Jiaya Jia[1,3]

[1] The Chinese University of Hong Kong
{rzwu, linhj, leojia}@cse.cuhk.edu.hk
[2] University of Hong Kong
xjqi@eee.hku.hk
[3] SmartMore

Abstract. Video propagation is a fundamental problem in video processing where guidance frame predictions are propagated to guide predictions of the target frame. Previous research mainly treats the previous adjacent frame as guidance, which, however, could make the propagation vulnerable to occlusion, large motion and inaccurate information in the previous adjacent frame. To tackle this challenge, we propose a memory selection network, which learns to select suitable guidance from all previous frames for effective and robust propagation. Experimental results on video object segmentation and video colorization tasks show that our method consistently improves performance and can robustly handle challenging scenarios in video propagation.

1 Introduction

Video propagation is a fundamental technique in video processing tasks, including video colorization [18, 46, 47], video semantic segmentation [27, 14], video object segmentation [29, 5, 20, 21, 16], to name a few. It aims at propagating information from an annotated or intermediate guidance frame to the entire video.

Prior work [29, 5, 20, 21, 16] mainly focused on propagating information in a frame-by-frame fashion as illustrated in Figure 1(a) where adjacent frames are utilized to update the target one. This propagation pipeline is fragile due to accumulation of errors, since inaccurate predictions in previous frames inevitably influence target frame prediction. The influence is magnified especially when the target object disappears or is misclassified.

To address the error accumulation caused by frame-by-frame propagation, one feasible solution is to utilize the information from all previous frames to propagate them to the current one, as illustrated in Figure 1(b). Albeit reasonable, these frames contain a lot of redundant and cluttered information, and the problem becomes more serious as the number of previous frames increases. Thus, selecting the best frame for propagating information effectively and robustly in videos is a critical issue.

⋆ Equal Contribution.

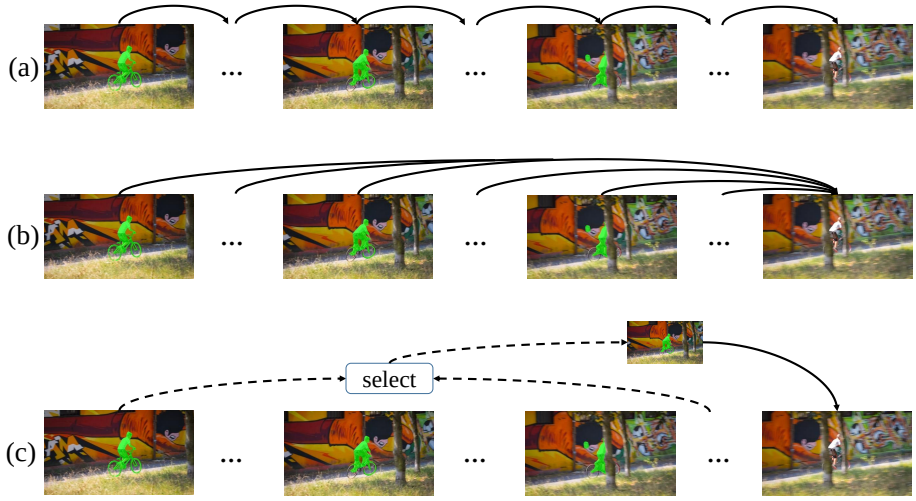

Fig. 1. Illustration of different strategies for propagating segmentation masks. (a) Frame-by-frame propagation. (b) Propagation with all previous frames. (c) Our proposed strategy for selecting the proper guidance frames for propagation.

In this paper, we propose a memory selection network (MSN) to vastly benefit generic video propagation. To update information in the current frame, shown in Figure 1(c), our MSN selects the most informative frames from a memory pool, which caches all previous frames as features. We note that this seemingly simple idea produces promising results. It effectively reduces error accumulation while not affecting computation cost much.

Our selection network serves as a generic and efficient component for video propagation to complement any propagation methods. Specifically, we apply our proposed memory selection network to different video propagation based approaches, including a classical temporal propagation network (TPN) built by us, and recent state-of-the-art propagation framework STM [28]. Their performance is boosted with incorporation of the memory selection network. Moreover, to further demonstrate the generality and usefulness, we conduct experiments on both video object segmentation and video colorization tasks. Our overall contributions are summarized below.

- We propose a memory selection network (MSN) to select suitable guidance frames for video propagation.
- MSN is generic to be integrated into any video propagation framework.
- Experimental results on video object segmentation and video colorization demonstrate that our approach boosts video propagation with limited computational cost.

2 Related Work

Video Propagation Propagation across image and video pixels is a common technique in various computer vision tasks, such as image/video colorization [18, 46, 47], matting [19], object segmentation [28, 39, 29, 14], and semantic segmentation [14]. Traditional priors for propagation are mainly optimization-based [18, 19], which minimize the energy function on a graph. In addition, filtering-based approaches [10, 32] propagate information using image or video filters, faster than the optimization-based method.

Recently, several methods model spatial or temporal pixel-pixel relationship with convolutional neural networks. Jampani et al. [14] used bilateral CNN to model the relationship between neighborhood pixels. Liu et al. [23, 24] developed an affinity map for pixel propagation with CNN. In addition, there are a lot of object-level propagation approaches [28, 39, 14, 29, 21, 4] using deep neural networks specifically designed for video object segmentation.

Most of the above propagation approaches treat adjacent previous frame as the guidance for propagation, allowing the system to easily accumulate errors through different propagation steps. Oh et al. [28] utilizes information from multiple previous frames and adaptively fuses them for propagation to the target.

In this paper, we design a generic module to select suitable frames for video propagation. It can be seamlessly inserted into these propagation approaches to improve stability, robustness and quality.

Semi-Supervised Video Object Segmentation Semi-supervised object segmentation refers to the problem of segmenting all corresponding objects annotated in the first frame. A group of frameworks were proposed to tackle this problem [2, 29, 28, 35, 6, 39, 40, 3, 7, 9]. Some of them [2, 29, 21] rely on the online learning technique, which requires time-consuming fine-tuning on the annotated frame for each testing sequence.

Among these approaches, one major stream contains propagation-based methods. MaskTrack [29] provides a classical propagation baseline method using the last frame mask or optical flow as guidance. Many following methods [39, 20, 16, 43, 28] are based on it and improve it with more components or better strategies. LucidTracker [16] incorporated additional data augmentation during online training. Li et al. [20] fixed long-term propagation errors by introducing a re-identification module to complement frame-by-frame propagation. The reference image is introduced as guidance for better propagation in the work of [43, 39]. The network design is also improved correspondingly. These approaches utilize the previous adjacent frame for propagation, which makes the system easily fail in long-term propagation. STM [28] utilizes more previous frames in an effective way. Based on these propagation approaches, we propose a selection strategy for the guidance frame to improve performance from another perspective.

For high-quality long-term propagation, ConvGRU or ConvLSTM structures [41, 34] were utilized to build an implicit memory module for long-term propagation. Such approaches may suffer from memory and optimization issues when

capturing long-range dependency during the training stage. Our method differs from these RNN-based methods in that we build an external memory pool to select the appropriate guidance frame without memory constraints.

Apparently similar work to ours is BubbleNet [9] since both BubbleNet and our work design an additional network to help boost performance. The difference is also clear and fundamental: the BubbleNet network aims to find the best frame to be annotated by a human before applying any propagation methods, while our work determines which of the previous predictions would be most helpful for prediction of the current frame.

Video Colorization Video colorization can also be addressed using video propagation approaches. Interactive colorization [44] propagates annotated strokes spatially across frames. The propagation procedure is guided by the matting Laplacian matrix and manually defined similarities. CNN-based methods [46, 47] achieved colorization with fully-automatic or sparsely annotated color. Recently, Liu et al. [24] proposed a switchable temporal propagation network to colorize all frames in a video using a few color key-frames. Additionally, methods of [14, 38] colorize the video sequence with the annotated first frame. To propagate annotated color information to the whole video, VPN [14] utilized a bilateral space to retrieve the pixel color and Vondrick et al. [38] leveraged pixel embedding for soft aggregation.

3 Proposed Method

3.1 Overview

We propose a generic memory selection network to select the appropriate guidance frame for general video propagation. In the following, we use the video object segmentation task as an example to illustrate our approach.

Formulation We denote a video sequence with T frames as $\{x_t | t \in [0, T-1]\}$, where x_t refers to the raw frame at time step t. Given the annotation information y_0 of the first frame x_0, the goal of video propagation is to propagate the information to the whole video, i.e. to produce $\{y_t | t \in [1, T-1]\}$ from time step 1 to $T-1$ via a propagation module $\mathcal{P}$. For each target frame x_t, $\mathcal{P}$ utilizes the guidance image x_g, and the corresponding prediction result or annotation y_g, to obtain y_t. This can be formulated as

$$y_t = \mathcal{P}(x_t, x_g, y_g). \tag{1}$$

Previous frame-by-frame propagation is derived as $y_t = P(x_t, x_{t-1}, y_{t-1})$, where x_{t-1} serves as the guidance frame for frame x_t. In contrast, our approach aims to select the suitable guidance frame $x_g \in \{x_0, x_1, ..., x_{t-1}\}$ for propagation.

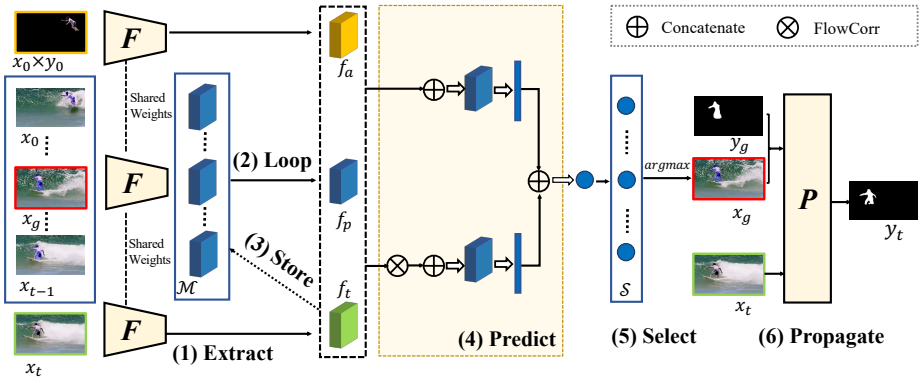

Fig. 2. Illustration of our framework. For each frame x_t in a video sequence, we first (1) **extract** feature f_t by a feature extractor $\mathcal{F}$, then we (2) **loop** for extracted features f_p in the memory pool $\mathcal{M}$, which is constructed from previous frames. At the same time, we also (3) **store** f_t back into $\mathcal{M}$ for later frames. For all f_p in $\mathcal{M}$, we (4) **predict** the selection score for each with the input of f_a, f_p and f_t, where f_a is the feature of the first frame masked with annotated objects. We then (5) **select** the frame x_g with the highest score $\mathcal{S}$ as guidance. Finally, x_g is utilized as the guidance frame for x_t to (6) **propagate** and obtain y_t. 'FlowCorr' is developed in FlowNet [13].

Workflow The overall workflow of our framework is shown in Figure 2. Our system builds a memory pool $\mathcal{M} = \{f_p | p = 0, 1, ..., T-1\}$ by sequentially caching features of previous frames, where f_p represents the extracted representation of frames x_p with the feature extractor network $\mathcal{F}$. We also extract feature of the first frame masked with the annotated objects as f_a for subsequent selection score prediction. To select a proper guidance frame for x_t, we extract its feature f_t at first and estimate the selection score for all features $\{f_p | p = 0, 1, ..., t-1\}$ cached in the memory pool via a light-weight selection network. It takes f_t, f_p and the feature of the annotated frame f_a as input, and outputs the corresponding selection scores. The frame with the highest score is selected as guidance for propagation, denoted as x_g. The propagation network $\mathcal{P}$ takes x_t, x_g and y_g as input to produce the final prediction y_t. f_t is cached back into $\mathcal{M}$ for subsequent frames.

It is worth noting that the feature extraction step takes much more time than the selection score estimation step, since the former is accomplished by a complicated network (i.e. VGG16 [33]) while the latter only uses a light-weight selection network consisting of only a few convolutional layers. Thus construction of the memory pool saves a lot of time by eliminating the feature extraction step of previous frames.

3.2 Memory Pool Construction

Representation The memory pool $\mathcal{M}$ is a set of features $\{f_p\}$, where $f_p \in \mathbb{R}^{512 \times \frac{H}{32} \times \frac{W}{32}}$, where H and W are the original spatial sizes. f_p is extracted from

the corresponding frame $x_p \in \mathbb{R}^{H\times W\times 3}$. We use a 2-D feature map instead of a vector to represent the memory because the spatial information is important in dense video propagation, e.g. video object segmentation and video colorization.

Construction Pipeline To construct the memory pool $\mathcal{M}$, the feature extractor $\mathcal{F}$ first takes the first frame x_0 as input and obtains the feature f_0. Then f_0 is cached to initialize $\mathcal{M}$. Additionally, we need the feature f_a concerning only the annotated object to make the selection operation aware of the target object. To this end, f_a is extracted from the annotated object x_a, which is obtained from x_0, whose background is masked by the annotated mask y_0. For each time step of the video sequence, the extracted feature of the target frame f_t is also cached into $\mathcal{M}$, which guarantees the efficiency that each frame only needs to be processed once by $\mathcal{F}$.

3.3 Memory Selection Network

Observation In frame-by-frame propagation, we have $y_t = \mathcal{P}(x_t, x_{t-1}, y_{t-1})$. Thus we can empirically infer that error accumulation of y_t stems from the prediction quality of y_{t-1} (the first factor) and the similarity between x_t and x_{t-1} (the second factor). We conduct experiments on YouTube-VOS [42] validation set [4] to verify the effect of these two factors (described below). For clarity, l_t indicates the ground-truth label of the t^{th} frame and $\mathrm{IoU}(\cdot,\cdot)$ indicates the intersection over union between two masks in the following description.

The influence of the prediction quality of y_{t-1} is illustrated in Figure 3(a). $\mathrm{IoU}(y_t, l_t)$ and $\mathrm{IoU}(y_{t-1}, l_{t-1})$ indicate the prediction quality of the last and target frames. As shown in Figure 3 (a), prediction quality of the $(t-1)^{th}$ and t^{th} frames is positively related, i.e. low quality y_{t-1} degrades y_t.

As for the other factor, the relation of y_t and similarity between x_t and x_{t-1} are plotted in Figure 3(b), where we use $\mathrm{IoU}(l_t, l_{t-1\to t})$ to represent similarity between the two frames, where $l_{t-1\to t}$ denotes the label warped from previous frame using optical flow [8]. It clearly draws the conclusion that the high similarity between x_t and x_{t-1} generally improves the propagation result y_t.

Selection Criterion According to the observations above, error accumulation in the frame-by-frame propagation pipeline is mainly influenced by two factors, i.e., the prediction quality of the guidance frame and its similarity with the target frame. Intuitively, if segmentation of the previous frame prediction is erroneous, the inaccurate information can be propagated to the target frame and accumulate dramatically across frames. Moreover, frames with high similarity reduce errors in the propagation stage and can help robust propagation.

Our selection network is designed to capture the above two factors. First, since the annotated frame is manually labeled by humans and is free of network

[4] YouTube-VOS online server returns a TEXT file containing the per frame IoU for each submission.

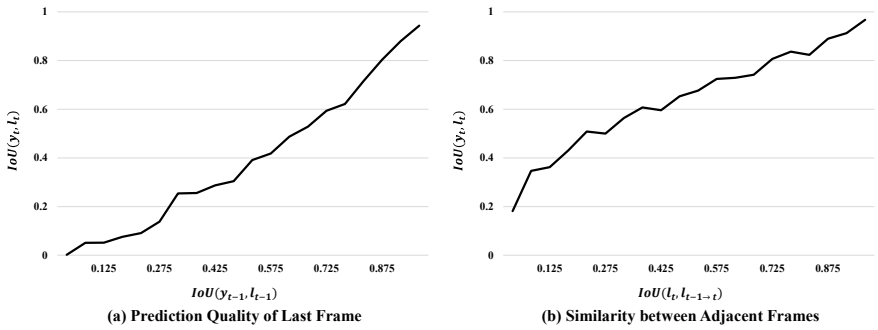

Fig. 3. Influence of two factors regarding error accumulation. Curves are plotted using TPN (Sec. 3.4).

prediction error, we adopt the feature map of the annotated frame f_a to help model the prediction quality of the guidance frame. We combine features from the annotated frame f_a and the guidance frame f_p with the FlowCorr operation $\text{FlowCorr}(f_p, f_a)$. The "FlowCorr" operation is developed in [13], which combines two feature maps by calculating the feature similarity between pixels. Next, to model similarity between f_p and target f_t, we further adopt the FlowCorr operation to combine their representation as $\text{FlowCorr}(f_p, f_t)$. The FlowCorr operation does not require computation of optical flow and is thus efficient.

Selection Network Design The selection module is designed as a binary classification network with 'good' and 'bad' categories. Specifically, for each feature f_p in the memory pool $\mathcal{M}$, the goal of the selection network is to calculate a score to measure utility for selecting it as the guidance frame regarding the target frame t. To this end, the selection network takes f_p, the annotated object's feature f_a, and target feature f_t as input and outputs a selection score. The highest-score one is selected as the guidance frame.

Based on the two key factors above, our selection network shown in Figure 2 adopts a two-stream structure. First, we use $\text{FlowCorr}(f_p, f_a)$, $\text{FlowCorr}(f_p, f_t)$ and $\text{FlowCorr}(f_t, f_a)$ to measure the relationship between guidance and annotated frames, guidance and target frames, and target and annotated frames, respectively. Then, a two stream network separately takes concatenation of $\{\text{FlowCorr}(f_p, f_a), \text{FlowCorr}(f_p, f_t), \text{FlowCorr}(f_t, f_a)\}$, and concatenation of $\{f_p, f_a, f_t\}$ as input and produces two feature vectors, which are further concatenated followed by a fully connected layer to generate the selection score. The detailed network structure is included in the supplementary material.

The memory selection network is light-weighted with only several convolution layers and fully-connected layers. The selection process can also be parallel for acceleration.

3.4 Video Propagation Frameworks

We select several video propagation based frameworks as baselines to verify the effectiveness of MSN.

Temporal Propagation Network (TPN) To verify the effectiveness of our proposed memory selection network, we build a classical temporal propagation network (TPN) as one baseline model. The design of TPN is similar to existing propagation-based frameworks [29, 21, 16]. It takes target frame x_t, selected guidance frame x_g, and corresponding predicted or annotated label y_g as input, and outputs the prediction label for the target frame y_t (i.e., Eq. (1)).

TPN consists of an appearance branch and a motion branch. The appearance branch takes x_t and y_g as input, while the motion branch takes optical flow $O_{t\to g}$ (between x_t and x_g) and y_g as input. Their output is further concatenated to obtain the final result. The detailed structure of TPN is described in our supplementary material. For each frame x_t, MSN selects proper x_g for TPN as input.

STM [28] STM is a state-of-the-art semi-supervised video object segmentation network. It is composed of three modules: 1) memory encoder, 2) query encoder, and 3) query decoder. The memory encoder encodes previous masks as well as corresponding frames into the memorized features. The target frame is encoded by the query encoder into a new feature and is further fused with the propagated memorized features. The fused feature is utilized to decode the mask for the target frame. We also incorporate MSN into STM by selecting the suitable memorized feature for propagating mask information.

3.5 Training Pipeline

Two Stage Training Since the $argmax$ operation is non-differentiable, the whole system adopts two-stage training. In the first stage, different video propagation frameworks are trained to converge. For training TPN, we adopt IoU loss [22] for video object segmentation and $\mathcal{L}_1$ regression loss for video colorization. For STM, we adopt their official pre-trained model in our experiments.

In the second stage, video propagation networks are fixed. They are used to generate the training samples to train the memory selection network. MSN is a binary classification network to estimate the quality of the guidance frame for the current frame in propagation. We adopt binary cross-entropy loss for MSN training. The method to generate positive ('good' guidance frames) and negative ('bad' guidance frames) training samples is elaborated below.

Generating Training Samples for MSN In the process of generating training samples for MSN, for each frame x_t in training sequences, we utilize the trained video propagation networks to propagate all previous frames $\{x_p|p = 0, 1, ..., t-1\}$ to x_t to obtain $t-1$ propagation results, denoted as $\{y_{p\to t}|p =$

$0, 1, ..., t-1\}$. With the label l_t of current frame, we obtain the IoU score of propagation results, i.e., $\{\text{IoU}(y_{p\to t}, l_t)|p = 0, 1, ..., t-1]\}$.

To split $\{x_p|p = 0, 1, ..., t-1\}$ into positive and negative samples, we first compute the highest IoU score IoU_{max} and the lowest score IoU_{min} among all these frames, and we set two thresholds σ_{pos} and σ_{neg} as hyper-parameters. The samples with IoU score in $[\text{IoU}_{max} - \text{IoU}_{max} * \sigma_{pos}, \text{IoU}_{max}]$ are split into positive samples, while those with score in $[\text{IoU}_{min}, \text{IoU}_{min} + \text{IoU}_{min} * \sigma_{neg}]$ are regarded as negative ones. The frames not belonging to either positive or negative samples are abandoned to avoid harming the classifier. These positive and negative samples are then used to train our memory selection network with binary cross-entropy loss. In our experiments, we empirically set the positive and negative thresholds as $0.05(\sigma_{pose})$ and $0.15(\sigma_{neg})$ respectively.

Implementation Details We use Adam [17] stochastic optimization, with the initial learning rate as 1e-5 and polynomial learning policy. The input image is resized to a fixed-size 640×320. TPN/MSN are trained on YouTube-VOS for 30/6 epochs and fine-tuned on DAVIS for 50/10 epochs. TPN is trained by randomly sampling two frames in a video as the guidance and target frames.

4 Experiments

We evaluate our proposed memory selection network on two different video propagation tasks: video object segmentation and grayscale video colorization. We focus on their semi-supervised setting where only the first frame is annotated with segmented mask or color. For the video object segmentation task, we evaluate our method on YouTube-VOS [42], and DAVIS 2016 and 2017 datasets [30, 31]. As for the video colorization dataset, following the work of [14], we conduct experiments on DAVIS 2016 dataset for evaluation.

The performance of the video object segmentation task is measured by region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ defined in [30]. Besides, For YouTube-VOS validation dataset, since there are 'seen' and 'unseen' categories, we provide $\mathcal{J}$ seen, $\mathcal{F}$ seen, $\mathcal{J}$ unseen and $\mathcal{F}$ unseen as corresponding metrics and *Overall* refers to the average score of them.

For grayscale video colorization, we evaluate the results with PSNR score and $\mathcal{L}_1$ distance between the generated results and its corresponding ground-truth.

4.1 Comparison with State-of-the-arts

Video Object Segmentation (VOS)

YouTube-VOS Dataset YouTube-VOS [42] dataset is the largest video object segmentation dataset with diverse objects, which contains 3471 training videos and 474 validation ones. The validation videos contain totally 91 object categories, with 65 seen categories and 26 unseen ones.

Table 1. Results of Video Object Segmentation on YouTube-VOS validation set. 'OL' denotes online training. '*' denotes using pre-trained weights on DAIVS Dataset [30, 31]. For all propagation based methods, we consider one-frame propagation.

Methods	Seen		Unseen			
	$\mathcal{J}(\%)$	$\mathcal{F}(\%)$	$\mathcal{J}(\%)$	$\mathcal{F}(\%)$	$Overall(\%)$	OL
OSVOS [2]	59.8	60.5	54.2	60.7	58.8	✓
MaskTrack [29]	59.9	59.5	45.0	47.9	53.1	✓
OnAVOS [37]	60.1	62.7	46.6	51.4	55.2	✓
S2S [41]	71.0	70.0	55.5	61.2	64.4	✓
PReMVOS [25]	**71.4**	**75.9**	**56.5**	**63.7**	**66.9**	✓
OSMN [43]	60.0	60.1	40.6	44.0	51.2	
RVOS [35]	63.6	45.5	**67.2**	51.0	56.8	
DMM [45]	60.3	63.5	50.6	57.4	58.0	
RGMP [39]	59.5	-	45.2	-	53.8	
A-GAME [15]	66.9	-	61.2	-	66.0	
TPN	64.0	65.9	57.0	65.4	63.0	
TPN + MSN	65.7	68.0	58.0	66.3	64.5 (+1.5)	
*STM-1	71.1	74.4	64.0	69.7	69.9	
*STM-1 + MSN	**72.4**	**75.2**	65.4	**71.4**	**71.1** (+1.2)	

We compare our method with state-of-the-art methods. The quantitative results are presented in Table 1. We utilize STM with only one previous frame for propagation as baseline, denoted as 'STM-1', and we apply MSN to select one frame to replace the previous frame ('STM-1 + MSN'). We provide results with TPN and STM-1 as baseline video propagation frameworks, and incorporate our memory selection module (MSN) into them as 'TPN + MSN' and 'STM-1 + MSN'. For both video propagation networks, we achieve consistent improvement, in terms of *Overall* score, of 1.5% and 1.2% respectively. We note since the pre-trained model of STM on YouTube-VOS dataset is not provided, we here adopt their pre-trained model on DAVIS for inference.

DAVIS 2016 and 2017 Datasets We further conduct experiments on DAVIS 2016 and 2017 datasets. DAVIS-2016 [30] is a popular single object segmentation benchmark, consisting of 30 training and 20 validation videos. DAVIS-2017 [31] is an extended version of DAVIS-2016 with multiple objects in a video sequence, consisting of 60 training and 30 validation videos.

We evaluate MSN with two baselines of TPN and STM [28] on the validation sets. Our memory selection module consistently benefits the baseline methods by choosing one suitable reference frame on both single-object- and multi-object-segmentation. MSN improves the baselines by 0.6% to 1.6% on both datasets, proving its effectiveness.

Visual Quality Results A selected sequence is visualized in Figure 4. For the results in TPN, the prediction error in the segmented mask accumulates and

Table 2. Comparison of video object segmentation methods on DAVIS 2016 and 2017 validation sets, where 'OL' indicates online learning techniques. For all propagation based methods, we consider one-frame propagation.

Methods	DAVIS-2016		DAVIS-2017		Runtime (s)	OL
	$\mathcal{J}(\%)$	$\mathcal{F}(\%)$	$\mathcal{J}(\%)$	$\mathcal{F}(\%)$		
OSVOS [2]	79.8	80.6	56.6	63.9	10	✓
PReMVOS [25]	84.9	**88.6**	**73.9**	**81.7**	-	✓
OSVOS-S [26]	85.6	86.4	64.7	71.3	4.5	✓
OnAVOS [37]	**86.1**	84.9	64.5	71.2	13	✓
CINM [1]	83.4	85.0	67.2	74.2	-	✓
MaskRNN [11]	80.7	80.9	60.5	-	-	✓
FAVOS [4]	82.4	79.5	54.6	61.8	**1.8**	✓
OSMN [43]	74.0	-	52.5	57.1	0.14	
VidMatch [12]	81.0	-	56.5	68.2	0.32	
FEELVOS [36]	81.1	82.2	69.1	74.0	0.51	
RGMP [39]	81.5	82.0	64.8	68.8	0.13	
A-GAME [15]	82.0	82.2	67.2	72.7	0.07	
DMM [45]	-	-	68.1	73.3	0.08	
TPN	75.8	74.2	58.9	62.7	0.17	
TPN+MSN	76.8	74.6	59.5	63.3	0.21	
STM-1 [28]	83.2	83.3	69.6	74.6	**0.06**	
STM-1 + MSN	**83.8**	**84.9**	**71.4**	**76.8**	0.10	

propagates along with the naive frame-by-frame propagation strategy. However, by selecting a proper guidance frame, we alleviate error accumulated and thus support high-quality long-term propagation.

Grayscale Video Colorization

Quantitative Results We also evaluate our proposed memory selection network on the grayscale video colorization task using the same training and inference strategies as the video object segmentation task. To quantify the effectiveness of our memory selection network, following VPN [14], we evaluate our algorithm on DAVIS-2016 dataset. For each video sequence, we take the first frame as the annotated color frame and propagate color to the rest of grayscale frames. PSNR and $\mathcal{L}_1$ between predicted target frame and ground-truth one in RGB color space are adopted as the evaluation metrics. Table 3 gives comparison among our framework and others. 'TPN + MSN' achieves the best performance in terms of both PSNR and $\mathcal{L}_1$, and MSN improves results a lot on this task.

Visual Quality Results Figure 5 shows visual results of a sample sequence. The color information reduces gradually by naive frame-by-frame propagation in TPN. TPN equipped with MSN preserves color information well since a better guidance frame is selected from the memory pool and propagated to each

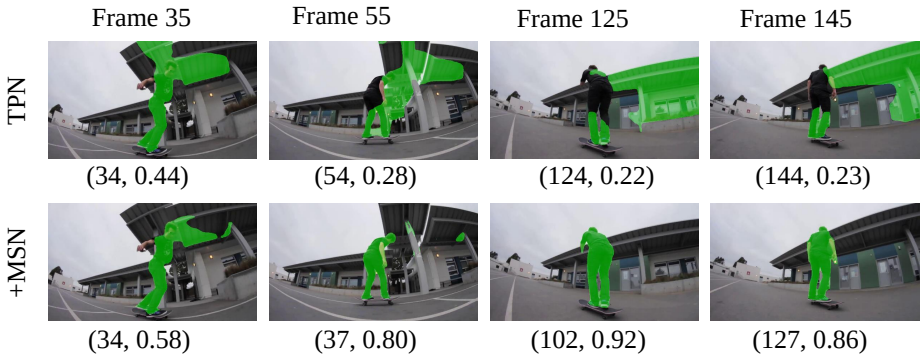

Fig. 4. Visualization of video object segmentation on YouTube-VOS validation set. For each target frame, $(\cdot,\cdot)$ denotes the index of selected guidance frame and the *Overall* score of propagated mask.

Table 3. Quantitative comparison of grayscale video colorization. ↑ means 'the higher the better'. ↓ means the opposite: 'the lower the better'.

Methods	PSNR ↑	$\mathcal{L}_1$ ↓	Runtime (s)
BNN-Identity [14]	27.89	13.51	0.29
VPN-Stage1 [14]	28.15	13.29	0.9
Levin et al [18]	27.11	-	19
TPN	28.25	11.06	**0.23**
TPN+MSN	**28.57**	**10.76**	0.27

target frame. The whole framework propagates color information much longer than the baseline propagation network, which greatly helps colorization for its final quality.

4.2 Ablation Study

Comparison of Selection Strategies In this section, we explore whether our designed selection network can be replaced by other simpler selection strategies.

- *VGG_select* : The guidance frame is selected by comparing its feature space distance with the target frame. The feature is extracted from pre-trained VGG [33] without fine-tuning.
- *VGG_mask_select* : To compare the distance, VGG feature distance of both the guidance frame and masked guidance frame with VGG feature of the target frame are separately computed and then added up.
- *Time step gap k* : The guidance frame is selected by a fixed time-step gap k with the target frame. For each frame x_t, the prediction is calculated by $y_t = \mathcal{P}(x_t, x_{max(0,t-k)}, y_{max(0,t-k)})$.

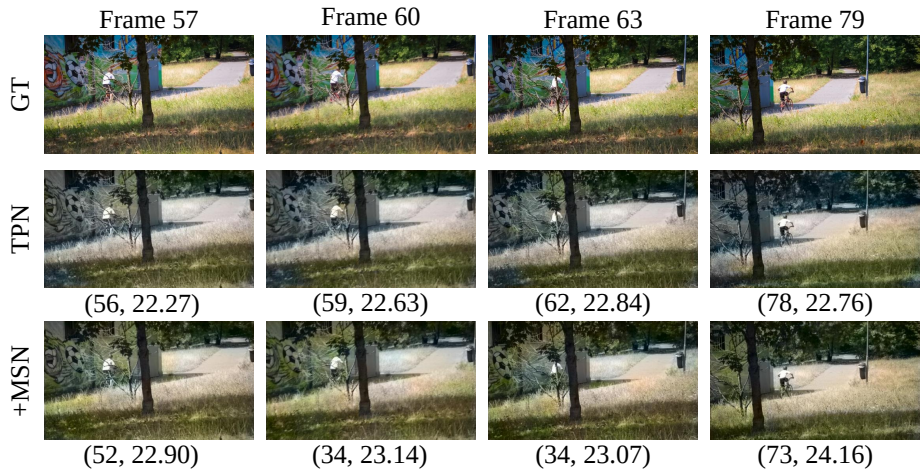

Fig. 5. Visualization of grayscale video colorization. $(\cdot,\cdot)$ below each frame denotes the index of selected guidance frame and the corresponding PSNR score with regard to ground-truth color image.

We test the above selection strategies as well as our trained memory selection network for video object segmentation on YouTube-VOS validation set. The performance is demonstrated in Table 4. Simply selecting the most similar guidance frames in VGG feature space is insufficient for propagation since pretrained VGG features are not aware of the propagation quality between frames. Moreover, the performance of simply selecting the frames with fixed time step gaps can be greatly erroneous since two frames far away may be significantly different in appearance. They increase difficulty of generating accurate motion information.

Table 4. Performance of different selection strategies on YouTube-VOS validation set. '*Overall*' metric defined in [42] measures the performance of different strategies.

Selection strategies	*Overall*
VGG_select	63.03
VGG_mask_select	63.08
Time step gap 1	63.04
Time step gap 5	63.13
Time step gap 10	59.67
Time step gap 20	55.61
MSN	**64.5**

Table 5. Performance and runtime for ensemble strategies. 'TPN-K' indicates ensembling predictions from the last K frames. '+MSN-K' indicates that the ensembled predictions are selected with the K highest selection scores.

Ensembles	*Overall*	Runtime
TPN-1	63.04	**0.09**
+MSN-1	64.54	0.13
TPN-3	65.27	0.29
+MSN-3	65.8	0.33
TPN-5	65.6	0.49
+MSN-5	**66.1**	0.53

Prediction Ensemble Ensemble is an important means to improve propagation accuracy in the inference stage. In frame-by-frame propagation, the predictions from the last K frames are ensembled to produce the t^{th} frame prediction. It is intriguing to investigate how to ensemble the predictions of selected guidance frames. Since MSN is trained as a binary classifier, the prediction score can be considered as the confidence of 'positive' for a guidance frame.

To ensemble of propagation for the memory selection network, the frames in the selection pool are ranked according to the scores obtained by the selection network. The top-K highest scoring frames are ensembled for the t^{th} frame. We conduct experiments on K and test its performance and runtime. As shown in Table 5, ensembling multiple guidance frames consistently increases the accuracy on different K.

Oracle Results We investigate the potential of MSN by applying ground-truth labels to select guidance frames. Specifically, for each target frame t with ground-truth label l_t, we compute the propagation results from all preceding frames, represented as $\{y_{p\to t}|p \in [0, t-1]\}$. The propagation mask with the highest accuracy $\text{IoU}(y_{p\to t}, l_t)$ is selected as the prediction mask for the t^{th} frame. As illustrated in Table 6, 'Oracle-MSN' achieves much better results than 'TPN' and 'TPN+MSN'. The results demonstrate that there is still much space to improve memory selection results.

Table 6. Oracle results in video object segmentation. '+MSN-Oracle' denotes selecting the guidance frame with the highest propagation accuracy. We report *Overall* and $\mathcal{J}$ scores for YouTube-VOS and DAVIS-2016, respectively.

Method	YouTube-VOS	DAVIS-2016
TPN	63.0	75.7
+MSN	64.5	76.4
+MSN-Oracle	75.3	82.4

5 Conclusion

We have presented a memory selection network for the robust video propagation by dynamically selecting the guidance frame to update information about the target frame. The memory selection network can select suitable guidance frames based on the quality of the guidance frame and its relationship with the target frame. Experimental results on video object segmentation and video colorization demonstrate that our method improves robustness of video propagation consistently.

References

1. Bao, L., Wu, B., Liu, W.: Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In: CVPR (2018)
2. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017)
3. Chai, Y.: Patchwork: A patch-wise attention network for efficient object detection and segmentation in video streams. In: ICCV (2019)
4. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: CVPR (2018)
5. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV (2017)
6. Ci, H., Wang, C., Wang, Y.: Video object segmentation by learning location-sensitive embeddings. In: ECCV (2018)
7. Duarte, K., Rawat, Y.S., Shah, M.: Capsulevos: Semi-supervised video object segmentation using capsule routing. In: ICCV (2019)
8. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis (2003)
9. Griffin, B.A., Corso, J.J.: Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In: CVPR (2019)
10. He, K., Sun, J., Tang, X.: Guided image filtering. TPAMI (2013)
11. Hu, Y.T., Huang, J.B., Schwing, A.: Maskrnn: Instance level video object segmentation. In: NeurIPS (2017)
12. Hu, Y.T., Huang, J.B., Schwing, A.G.: Videomatch: Matching based video object segmentation. In: ECCV (2018)
13. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017)
14. Jampani, V., Gadde, R., Gehler, P.V.: Video propagation networks. In: CVPR (2017)
15. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: CVPR (2019)
16. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for multiple object tracking. arXiv:1703.09554 (2017)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: TOG (2004)
19. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. TPAMI (2008)
20. Li, X., Qi, Y., Wang, Z., Chen, K., Liu, Z., Shi, J., Luo, P., Loy, C.C., Tang, X., Khoreva, A., et al.: Video object segmentation with re-identification. In: The 2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops (2017)
21. Li, X., Change Loy, C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In: ECCV (2018)
22. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: CVPR (2018)
23. Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.H., Kautz, J.: Learning affinity via spatial propagation networks. In: NeurIPS (2017)
24. Liu, S., Zhong, G., De Mello, S., Gu, J., Yang, M.H., Kautz, J.: Switchable temporal propagation network. arXiv:1804.08758 (2018)

25. Luiten, J., Voigtlaender, P., Leibe, B.: Premvos: Proposal-generation, refinement and merging for video object segmentation. In: ACCV (2018)
26. Maninis, K.K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: Video object segmentation without temporal information. TPAMI (2018)
27. Miksik, O., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient temporal consistency for streaming video scene analysis. In: ICRA (2013)
28. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019)
29. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: CVPR (2017)
30. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
31. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
32. Rick Chang, J.H., Frank Wang, Y.C.: Propagated image filtering. In: CVPR (2015)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
34. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV (2017)
35. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: Rvos: End-to-end recurrent network for video object segmentation. In: CVPR (2019)
36. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: CVPR (2019)
37. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. arXiv:1706.09364 (2017)
38. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: ECCV (2018)
39. Wug Oh, S., Lee, J.Y., Sunkavalli, K., Joo Kim, S.: Fast video object segmentation by reference-guided mask propagation. In: CVPR (2018)
40. Xu, K., Wen, L., Li, G., Bo, L., Huang, Q.: Spatiotemporal cnn for video object segmentation. In: CVPR (2019)
41. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: ECCV (2018)
42. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv:1809.03327 (2018)
43. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: CVPR (2018)
44. Yatziv, L., Sapiro, G.: Fast image and video colorization using chrominance blending. TIP (2006)
45. Zeng, X., Liao, R., Gu, L., Xiong, Y., Fidler, S., Urtasun, R.: Dmm-net: Differentiable mask-matching network for video object segmentation. In: ICCV (2019)
46. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
47. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. TOG (2017)