

# A Novel Video Salient Object Detection Method via Semi-supervised Motion Quality Perception

Chenglizhao Chen<sup>1</sup> Jia Song<sup>1</sup> Chong Peng<sup>1\*</sup> Guodong Wang<sup>1</sup> Yuming Fang<sup>2</sup>  
<sup>1</sup>Qingdao University <sup>2</sup>Jiangxi University of Finance and Economics

**Abstract**—Previous video salient object detection (VSOD) approaches have mainly focused on designing fancy networks to achieve their performance improvements. However, with the slow-down in development of deep learning techniques recently, it may become more and more difficult to anticipate another breakthrough via fancy networks solely. To this end, this paper proposes a universal learning scheme to get a further 3% performance improvement for all state-of-the-art (SOTA) methods. The major highlight of our method is that we resort the “motion quality”—a brand new concept, to select a sub-group of video frames from the original testing set to construct a new training set. The selected frames in this new training set should all contain high-quality motions, in which the salient objects will have large probability to be successfully detected by the “target SOTA method”—the one we want to improve. Consequently, we can achieve a significant performance improvement by using this new training set to start a new round of network training. During this new round training, the VSOD results of the target SOTA method will be applied as the pseudo training objectives. Our novel learning scheme is simple yet effective, and its semi-supervised methodology may have large potential to inspire the VSOD community in the future.

**Index Terms**—Motion Quality Assessment; Video Salient Object Detection; Semi-supervised Learning.

## I. INTRODUCTION AND MOTIVATION

Different from images that comprise spatial information only, video data usually contain both spatial (appearances) and temporal (motions) information. To alleviate the computational burden, most of the video related applications [1], [2], [3], [4], [5], [6], [7] have adopted the video salient object detection (VSOD) approaches as the pre-processing tool to filter the less important video contents while highlighting the salient objects that attract our visual system most, aiming to strike the trade-off between efficiency and performance.

After entering the deep learning era, the state-of-the-art (SOTA) VSOD approaches have achieved steady performance improvements via various fancy networks, such as ConvLSTM [10] and 3D ConvNet [11]. However, with the slow-down in development of the deep learning techniques recently, we shouldn't anticipate for new breakthrough via fancy networks solely. For example, compared with the leading SOTA method in 2019 (i.e., MGA [12]), the performance improvement made by the most recent work in 2020 (i.e. PCSA [13]) is really marginal with a performance gap less than 1% averagely. This fact motivates us to wonder why wouldn't we develop a universal learning scheme, rather than using fancy networks, to get the SOTA performances further improved?

Given an off-the-shelf VSOD approach (we name it as the “target SOTA method”), this paper aims to improve its performance via a novel learning scheme, and we formulate our idea as following.

1) We shall select a sub-group of video frames from the original testing set to construct a new training set, and these selected frames are needed to be the ones that have been “successfully detected” by the target SOTA method.

2) Consequently, we will achieve a significant performance improvement by using this new training set to start a new round of network training, in which the VSODs of the corresponding SOTA method will be used as the training objectives (pseudo-GT).

So, without using any saliency ground truth (GT) of the original testing set, all that remains now is how can we know which frames will be successful detected by the target SOTA method in advance.

Our key idea is quite simple and straight-forward, which is inspired by a common phenomenon in the SOTA methods; i.e., for most of the SOTA VSOD methods, their performances usually vary from frame to frame, even though these frames belong to an identical video sequence sharing similar scenes. For example, as is shown in Fig. 1, the 1st row shows 10 consecutive frames with similar scenes containing a *worm* as the salient object; however, as is shown in the 3rd row, the VSOD results of the SOTA method (SSAV [9]) in the frame #17, #18, #21 and #24 are clearly better than other frames. The main reason is that the VSOD performance is determined by both spatial and temporal saliency clues. Though the spatial saliency clues are usually stable between consecutive video frames, the motion saliency clues may vary a lot due to the unpredictable nature of movements, not to mention other additional challenges induced by camera view angle changes. So, we propose a brand new concept—“motion quality”, to predict which video frames will have large probability to be successfully detected by the target SOTA approach.

For those clear motions (e.g., rigid movements) which can positively facilitate the VSOD task by separating salient objects from their non-salient surroundings nearby, we name it as the “high-quality motions”, and we call other cases as the “low-quality motions” accordingly, see Fig. 2.

In most cases, we believe that those video frames containing “high-quality motions” should be selected into our new training set. To predict motion quality in advance, we advocate a semi-supervised scheme to train our motion quality perception module (MQPM) within a frame-wise manner, see the Fig. 3-C and it will be detailed in Sec. III-B. As one of the

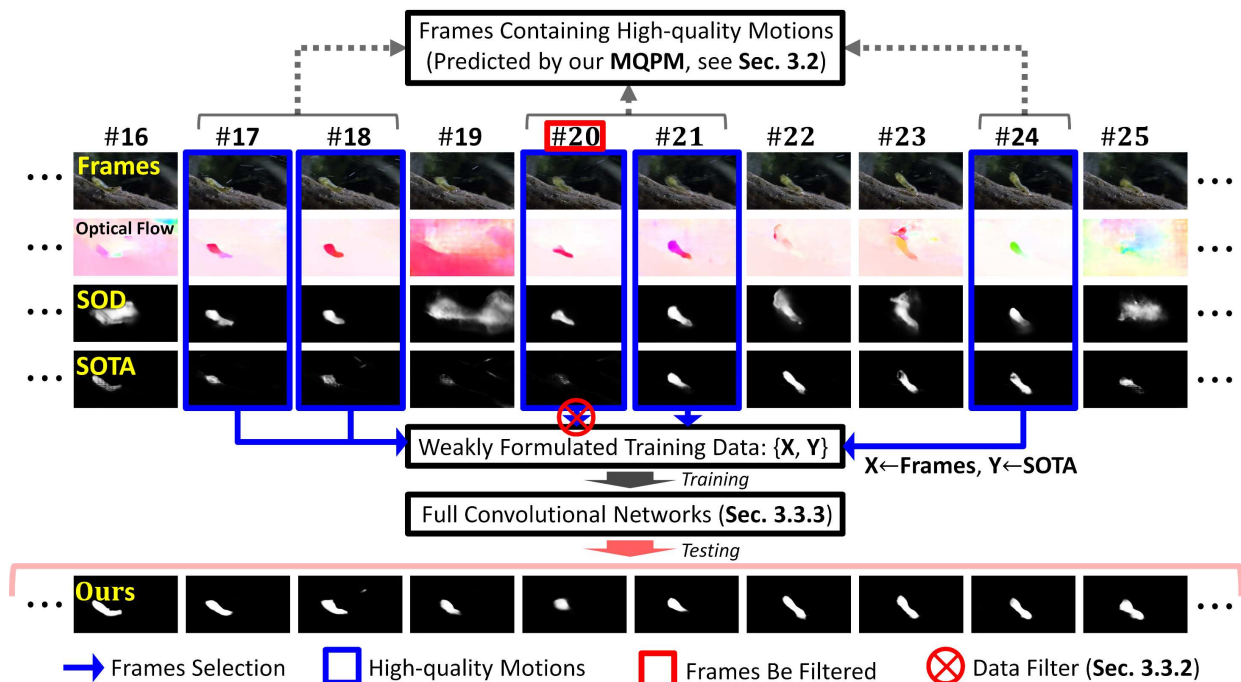


Fig. 1. The key motivation of our method is to select a sub group video frames from the original testing set to construct a new training set, and these selected frames should have high-quality motions (by our MQPM) and their VSOD provided by the target SOTA method will be used as the pseudo-GT to start a new round training which will improve the target SOTA method significantly. **SOD**: the salient object detection results by feeding the optical flow data into the pre-trained image salient object detection model (we choose CPD [8] here, see Eq. 1); **SOTA**: the VSOD of the target SOTA method (we take the SSAV [9] for example) which we aim to improve its performance, and it can be any other SOTA methods; **Ours**: the final VSOD results after using our novel learning scheme, of which the overall performance have outperformed the SOTA results, significantly.



Fig. 2. Motion quality demonstrations, where the high-quality motions can usually separate salient objects from their non-salient surroundings nearby, while the low-quality motions cannot achieve this.

key components in our learning framework, the MQPM takes motion patterns (sensed by optical flow) as input, and then it makes a binary decision regarding whether or not the given frame contains high-quality motions. Meanwhile, in the case of a video frame has some high-quality motions, the MQPM will also provide the corresponding spatial locations of these high-quality motions, and these spatial locations will be used to facilitate the data filtering scheme (Sec. III-C2), another key component in our learning framework, to double-check if these motions really belong to the high-quality cases.

In summary, the main contributions of our method can be summarized as following four aspects:

- A semi-supervised learning scheme to conduct Motion Quality Perception (to the best of our knowledge, this is the first attempt to improve the VSOD performance from the motion quality perspective);
- A universal scheme to improve the performance of any

other SOTA methods (at least 3% performance improvement in general);

- Extensive quantitative validations and comparisons (almost all SOTA methods in recent 3 years over 5 largest datasets);
- Method source code and results are all publicly available at <https://github.com/qduOliver/MQP>, which will have large potential to benefit the VSOD community in the future.

## II. RELATED WORK

### A. Image Salient Object Detection

The main target of image saliency ([14], [15], [16]) is to fast locate the most eye-catching objects in a given image. In general, there are two typical methods for the image salient object detection (ISOD) task, including the full convolutional networks (FCNs) based methods and the multi-task learning (MTL) based methods, and we will briefly introduce several most representative methods regarding these two types.

1) *The FCNs based methods*: The key rationale of the FCNs based methods [17], [18], [8] is to utilize the multi-scale/multi-level contrast computation to sense saliency clues. In fact, different network layers usually show different saliency perception abilities, i.e., those deeper layers tend to preserve localization information solely, yet those shallower layers are mainly abundant in tiny details. Thus, Hou *et al.* [17] proposed to use short connections between different layers to achieve the multi-scale ISOD, in which the coarse localization information was introduced into the shallower layers, achieving a much improved performance. Similarly, Wang *et al.* [18] adopted

a top-down and bottom-up inference network, implementing step-by-step optimization via a cooperative and iterative feed-forward and feed-back strategy. Although these two most representative methods have achieved significant performance improvements, their network structures are generally too heavy. In contrast, Wu *et al.* [8] proposed a lightweight framework, which discarded those high-resolution deep features to speed up detection, of which the motivation is that those deep features in shallower layers usually contribute less to the overall performance yet at high computational costs.

2) *The MTL based methods*: The key rationale of the MTL based methods is to resort additional auxiliary information to boost the overall performance of the conventional single stream methods, in which such information frequently includes depths [19], image captions [20] and edge clues [21], [22], [23]. Zhu *et al.* [19] proposed to learn a switch map to adaptively fuse the RGB saliency clues with the depth saliency clues to formulate final ISOD result. Zhang *et al.* [20] leveraged the image captions to facilitate their newly proposed weakly supervised ISOD learning scheme, in which the key idea is to utilize the feature similarities between different caption categories to shrink the given problem domain. Qin *et al.* [22] proposed a novel edge related loss function to further refine the tiny details in the final ISOD maps. Similarly, Zhao *et al.* [23] combined the edge loss function with multi-level features to further improve the ISOD performance, in which the edge related saliency clues are treated as an explicit indicator to coarsely locate the salient objects.

## B. Video Salient Object Detection

1) *Conventional hand-crafted methods*: Different to the above mentioned ISOD methods, the video salient object detection (VSOD) is more challenge due to the newly available temporal information. Previous hand-crafted approaches [24], [25], [26], [27], [28] have widely adopted the low-level saliency clues, which were revealed individually from either spatial branch or temporal branch, to formulate their VSOD. To fuse spatial and temporal saliency clues, Wang *et al.* [24] resorted both the spatial edges and the temporal boundaries to facilitate the salient object localization. Guo *et al.* [25] designed a primitive approach to identify the salient object by ranking and selecting the salient proposals. Chen *et al.* [26] devised a bi-level learning strategy to model long-term spatial-temporal saliency consistency. Guo *et al.* [27] proposed a fast VSOD method by using the principal motion vectors to represent the corresponding motion patterns, and such motion message coupling with the color clues together will be fed into the multi-clue optimization framework to achieve the spatiotemporal VSOD.

2) *Deep-Learning based methods*: The development of convolutional neural networks (CNNs) has fulfilled the needs for performance improvement in the VSOD field. To date, since the spatial saliency can be measured via the off-the-shelf ISOD deep models, considerable researches have been paid to the measurement of temporal saliency within the deep learning framework, in which the current mainstream works can be categorized into two groups according to their network

structures [29], i.e., the single-stream network based methods and the bi-stream network based methods.

We will introduce the single-stream network based methods firstly. Le *et al.* [30] designed an end-to-end 3D network to directly learn spatiotemporal information. This 3D framework has added a refinement component at the end of its encoder-decoder backbone network, and its key rationale is to resort the semantical information of the deeper layers to refine its spatiotemporal saliency maps. Li *et al.* [31] developed a novel FCNs based network to conduct VSOD within a stage-wise manner which mainly consists of two main stages; i.e., the spatial saliency maps (using RGB information solely) will be computed in advance, and then those spatial saliency maps within consecutive video frames will be simultaneously fused as the spatiotemporal saliency maps. To enlarge the temporal sensing scope, Wang *et al.* [32] adopted the optical flow based correspondences to warp long-term information into the current video frame. Similarly, Song *et al.* [33] presented a novel scheme to sense the multi-scale spatiotemporal information, in which the key idea is to resort the bi-LSTM network to extract long-term temporal features. Meanwhile, this work has adopted the pyramid dilated convolutions to extract multi-scale spatial saliency features, which will latterly be fed into the above mentioned bi-LSTM network to achieve the long-term and multi-scale VSOD. Fan *et al.* [9] developed an attention-shift baseline and also released a large-scale saliency-shift-aware dataset for the VSOD problem.

Different from the single-stream networks with limited motion sensing ability [34], [35], the bi-stream networks [36], [37] are usually capable of sensing the motion clues explicitly, in which both the RGB frames and the optical flow maps are treated as the input of their two subbranches, individually. Then, both the spatial saliency clues and the temporal saliency clues will be computed respectively and latter be fused as the final VSOD results. Tokmakov *et al.* [38] proposed to feed the concatenated spatial and temporal deep features into the ConvLSTM network, aiming to strike an optimal balance between its temporal branch and spatial branch. Li *et al.* [12] exploited the motion message as attention to boost the overall performance of its spatial branch. Most recently, Gu *et al.* [13] learned the non-local motion dependencies across several frames, and then it followed the pyramid structure to capture the spatiotemporal saliency clues at various scales.

## III. PROPOSED APPROACH

### A. Method Overview

Given a pre-trained SOTA method (i.e., the target SOTA method), our key idea is to use a subgroup of testing frames with high-quality VSODs to train a novel appearance model, and this novel model will significantly outperform the target SOTA method eventually. To achieve it, our method mainly consists of three steps, and the detailed method overview can be found in Fig. 3.

- 1) Firstly, we weakly train a novel deep model, i.e., the Motion Quality Perception Module (MQPM, blue box).
- 2) Next, we use the MQPM to select a subgroup video frames (with high-quality motions) in testing set to formulate a new

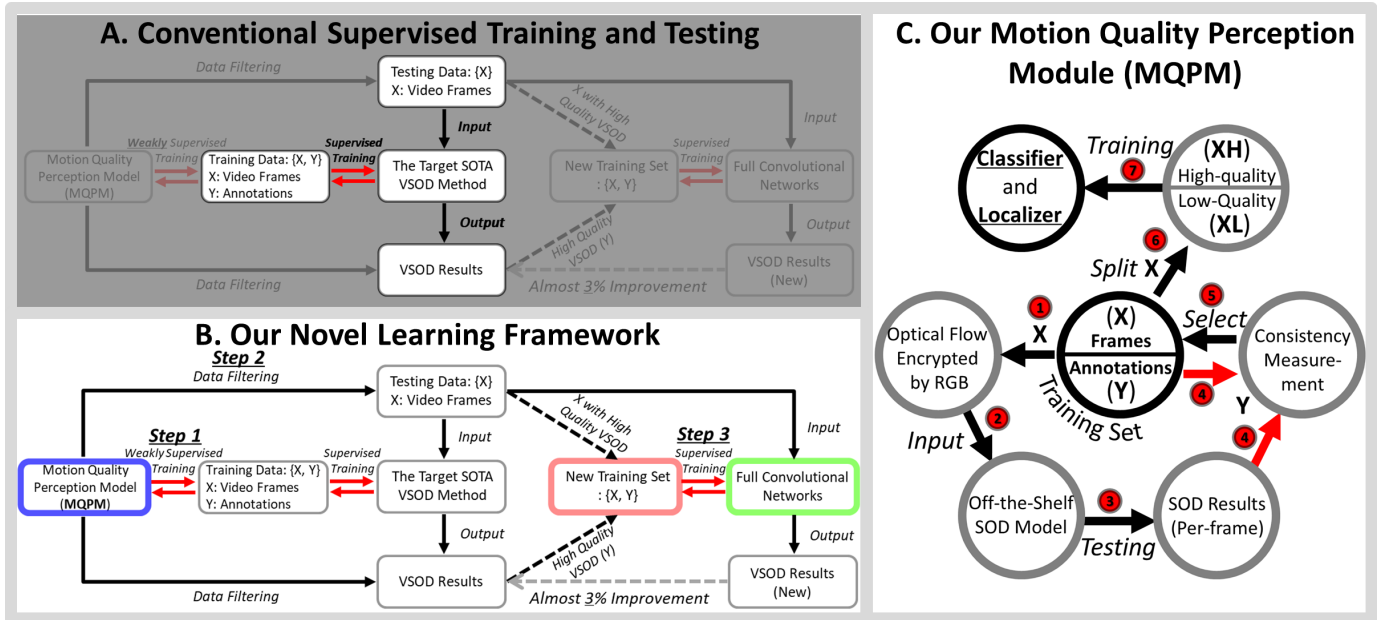


Fig. 3. The overall method pipeline. Our novel learning scheme can be applied to the conventional learning scheme (see subfigure-A), and it mainly consists of three steps which have been marked by different colors (red, green and blue) in the subfigure-B; The motion quality perception module is the most important component, and we have demonstrated its details in subfigure-C, where the marks from 1 to 7 respectively show the detailed dataflow.

training set (red box).

3) Finally, this new training set will be used to train a novel appearance model with much improved VSOD performance (green box).

### B. Motion Quality Perception Module

We demonstrate the detailed MQPM pipeline in Fig. 3-C, the ultimate goal of our approach is to provide a frame-wise binary prediction regarding whether or not the given frame contains high-quality motions. If yes, it will also provide the spatial locations of the high-quality motions.

To achieve our goal, we should initially divide the training instances (i.e., frames) of the original VSOD training set (i.e., Davis-TR [39]) into two groups, i.e., one includes frames with high-quality motions, and another one includes frames with low-quality motions only. Thus, the MQPM can be easily trained by using this partition.

Now the problem is how can we automatically achieve such motion-quality-aware partition in advance.

1) *Motion Quality Measurement*: As is shown in the 2nd row of Fig. 1, we have demonstrated the corresponding optical flow results (encrypted using RGB color) of some consecutive frames in a given video sequence (i.e., the “worm” sequence from the widely-used Davis set). Notice that these optical flow results are computed by using the off-the-shelf optical flow tool [40] to sense motions between two consecutive frames, in which the RGB colors at different pixels denote the estimated motion intensities and directions. It can be easily observed in Fig. 1 that the video frames with high-quality motions (e.g., the frame #18) usually share some distinct attributes in common, i.e., the optical flow values inside the salient object (i.e., the worm) will be totally different to the non-salient surroundings

nearby. Based on this, we propose a simple yet effective way to measure the motion quality score (MQS) as Eq. 1 with a quite straight forward rationale; i.e., the salient objects in those frames with high-quality motions will have large probability to be successfully detected by the off-the-shelf image salient object detection method, and these frames should be assigned with large MQSs.

$$MQS_i = f(\Theta(OF_i), GT_i), \quad (1)$$

where  $OF_i$  denotes the optical flow result of the  $i$ -th frame, and  $GT$  denotes the human well-annotated pixel-wise VSOD saliency ground truth;  $\Theta$  denotes a pre-trained image salient object detection deep model, which we choose the off-the-shelf CPD [8] due to its lightweight implementation;  $f$  denotes the consistency measurement between the SOD made by  $\Theta$  and the  $GT$ . In fact, there are various consistency measurements which are widely used to conduct quantitative evaluations, such as MAE [41], F-Measure [42] and S-Measure [43]. For simplicity, we choose the S-Measure as the consistency measurement  $f$  in Eq. 1. Notice that we have also tested other measurements, but the overall performance won’t change much, i.e., floating of two decimal places mostly.

2) *Training Set for MQPM*: To train our MQPM, we need to weakly assign binary labels for each frame in the Davis training set regarding whether it contains high-quality motions. Therefore, we use the motion quality scores (MQS, Eq. 1) as the key indicator to produce such labels ( $Label_i$ ) as following:

$$Label_i = \begin{cases} 0 & \text{if } MQS_i < \lambda \\ & \text{(i.e., XL in Fig. 3C)} \\ 1 & \text{otherwise} \\ & \text{(i.e., XH in Fig. 3C)} \end{cases}, \quad (2)$$

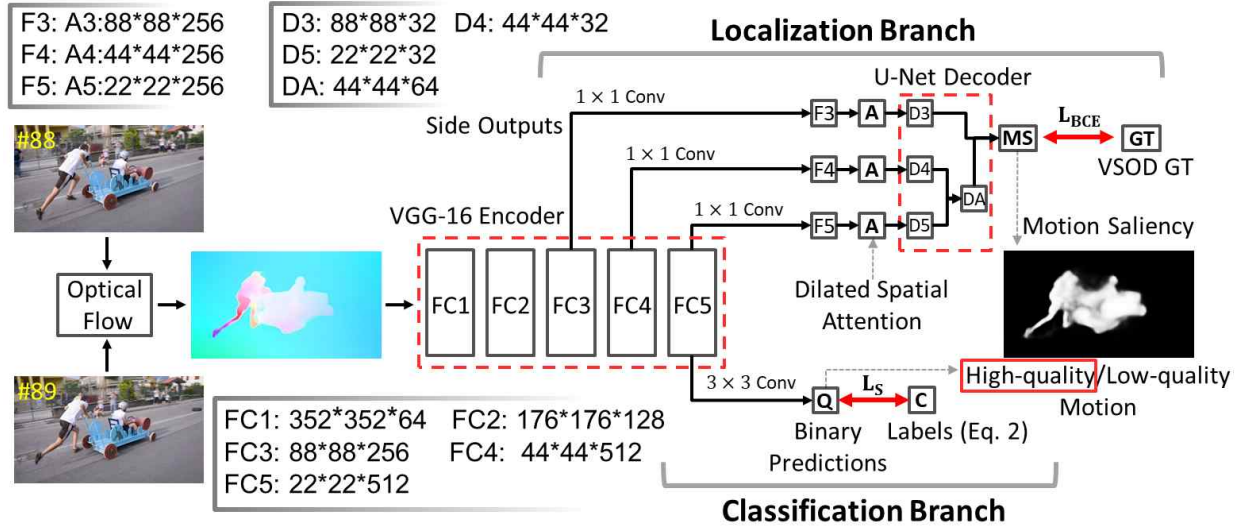


Fig. 4. The detailed network architecture of our motion quality perception module (MQPM). For simplicity, we have omitted all up-sampling/down-sampling operations.

where XH means high-quality optical flow frames, and XL denotes low-quality movements. When motion quality scores (MQS) is less than the threshold value  $\lambda$ , the label is assigned to 1. Otherwise, the label is assigned to 0, where  $\lambda$  is a pre-defined decision threshold. To ensure an optimal balance between positive-1 and negative-0 training instances, we iteratively update  $\lambda$  until the convergence via Eq. 3 and Eq. 4.

$$\omega = \frac{\int_{\lambda}^{\infty} \text{MQS} \cdot P(\text{MQS}) \cdot d(\text{MQS})}{\int_{\lambda}^{\infty} P(\text{MQS}) \cdot d(\text{MQS})}, \quad (3)$$

$$\lambda = (1 + \omega)/2, \quad (4)$$

where  $P(\text{MQS})$  is the probability distribution of MQS in the entire VSOD training set.

Thus far, we can formulate the training set as  $\{X_i, GT_i, Label_i\}$ , where  $X_i$  denotes the  $i$ -th video frame,  $GT$  is the original binary VSOD ground truth. Next, we will introduce how to train the MQPM by using this training set.

3) *MQPM Training*: We formulate our MQPM training as a multi-task procedure following the vanilla bi-stream structure, in which one stream aims the binary motion quality prediction (i.e., classification) and another stream conducts the pixel-wise motion saliency detection (i.e., localization).

As is shown in Fig. 4, the MQPM takes the RGB encrypted optical flow data as input, and its output comprises two parts: 1) motion saliency map; 2) motion quality prediction. The main network structure of MQPM comprises three components: one feature encoder (VGG-16 [44]) and two sub-branches with different loss functions.

The motion saliency branch takes the last three encoder layers as input. Next, each of these input will be fed into the widely-used multi-scale dilated attention module (with dilation factors ranging between  $\{2,4,6,8\}$ ) to filter those irrelevant features. Thus, the motion saliency map can be computed by applying the U-Net [45] decoder iteratively, in which the binary cross entropy loss ( $L_{BCE}$ ) is used. Meanwhile, the

classification branch only takes the last decoder layer as input. Thus, the total loss function can be represented as Eq. 5.

$$L_{total} = L_{BCE} + L_S, \quad (5)$$

where the binary cross entropy loss ( $L_{BCE}$ ) can be detailed as Eq. 6, and the  $L_S$  is a typical binary classification loss as shown in Eq. 7.

$$L_{BCE} = - \sum_i \sum_u GT_i(u) \times \log \left( MS_i(u) \right) - \sum_i \sum_u \left( 1 - GT_i(u) \right) \times \log \left( 1 - MS_i(u) \right), \quad (6)$$

where  $MS_i(u)$  denotes the predicted motion saliency value at the  $u$ -th pixel in the  $i$ -th frame;  $GT_i(u)$  represents ground truth value at the  $u$ -th pixel in the  $i$ -th frame; “ $\times$ ” is a conventional multiplication operation;  $\log$  is a typical logarithmic mathematical operation.

$$L_S = - \sum_i \left[ Label_i \times \log Q_i + (1 - Label_i) \times \log(1 - Q_i) \right], \quad (7)$$

where  $L_S$  is a logistic regression cost loss function;  $Q_i$  denote the confidence regarding the category predictions (i.e., high-quality/low-quality motions);  $Label_i$  is the previously determined motion quality label (Eq. 2).

### C. New Training Set For VSOD

1) *Initialization*: Thus far, the motion quality perception module (MQPM) has been trained, providing two vital information which can be used to improve the target SOTA method: 1) the binary motion quality prediction; 2) the motion saliency map.

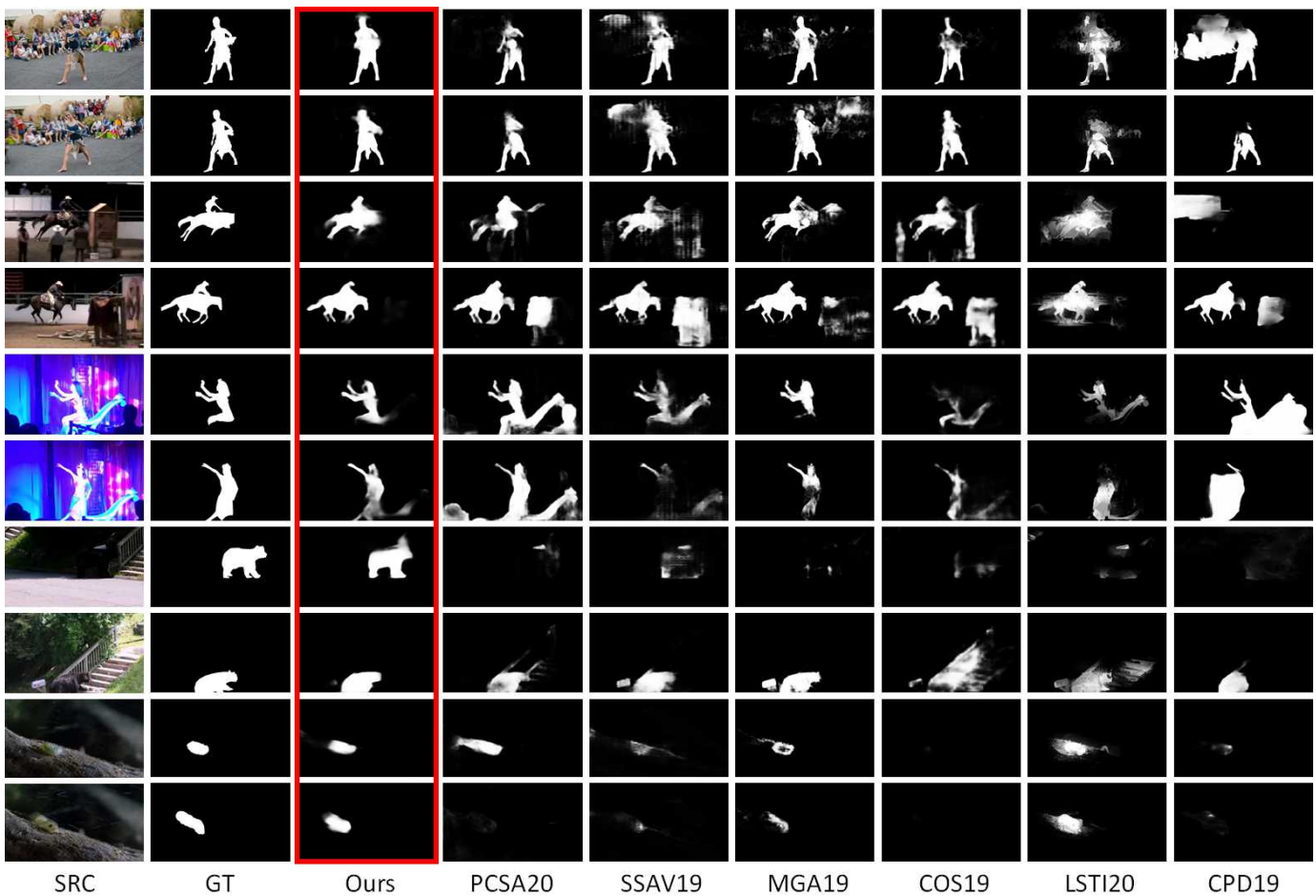


Fig. 5. Qualitative comparisons with the current SOTA methods. Due to the limited space, we only list six most representative ones here, including PCSA20 [13], SSAV19 [9], MGA19 [12], COS19 [46], LSTI20 [47] and CPD19 [8].

As we have mentioned before, the former one can be used as an explicit indicator to tell which frames in the VSOD testing set should be selected, while the latter will be used as a double-check to ensure the selected frames are really with high-quality motions which are capable of benefiting the VSOD training in practice. Here we will use both of these two to facilitate the construction of a new training set, which only comprises video frames containing high-quality motions. And this new training set will be used to start a new round of network training and improve the target SOTA method performance eventually.

For each frame in the VSOD testing set, we first compute its optical flow results frame-by-frame, and then feed these optical flow results into the well-trained MQPM, and thus those frames (i.e., the original frames rather than their optical flow results) which are predicted to have contained “High-quality Motions” will be directly pooled as the initial version of the new training set.

Next, for each training instance ( $\{X_i, Y_i\}$ ) in this new training set, it mainly consists of two components, including the original frame  $X$  and the corresponding VSOD result predicted by the target SOTA method (trained using both spatial and temporal information) as its training objective (i.e.,  $Y$ , see the pictorial demonstration in the red box of Fig. 3).

Also, it is worthy mentioning that we can not directly use

the motion saliency maps (i.e., the output of the localization branch in Fig. 4) as the training objectives. The main reason is that the motion saliency maps are usually with blur object boundaries (due to absent of spatial information), and thus the performance improvement may be severely limited if we directly apply the motion saliency maps as the pseudo-GT during this new round training, and the corresponding quantitative evidences can be found in Table. III.

2) *Data Filtering*: As we have mentioned before, our rationale is based on the assumption that the SOTA methods tend to exhibit high-quality VSOD over those frames with high-quality motions (see the quantitative proofs in Table. II). In fact, this assumption holds in most cases. However, there still exists exceptions occasionally.

As is shown in Fig. 1, our MQPM has predicted that the #20 frame has large probability of containing some high-quality motions, and the optical flow result of the #20 frame (in the 2nd row) is indeed capable of separating the salient object from its non-salient surroundings nearby, producing high-quality motion saliency map as well (in the 3rd row). However, the VSOD predicted by the target SOTA method (i.e., it can be any SOTA method, here, we simply choose the SSAV [9] for example) failed to completely detect the salient object, and it may degrade the overall performance if the new

TABLE I  
ABLATION STUDY REGRADING OUR DATA FILTERING STRATEGY, WHERE BASELINE DENOTES THE TARGET SOTA METHOD (I.E., S<sub>SSAV</sub>), SEE MORE DETAILS IN SEC. IV-E.

Dataset	Davis			Segv2			Visal			DAVSOD			VOS		
Metric	maxF	S-M	MAE	maxF	S-M	MAE	maxF	S-M	MAE	maxF	S-M	MAE	maxF	S-M	MAE
Baseline	0.861	0.893	0.028	0.801	0.851	0.023	0.939	0.943	0.020	0.603	0.724	0.092	0.742	0.819	0.073
T=1	0.892	<b>0.910</b>	0.019	0.826	0.873	0.020	0.934	0.939	0.018	0.686	0.764	0.076	0.755	0.820	0.067
T=1/2	0.890	0.908	0.018	0.824	0.866	0.019	0.934	0.935	0.018	0.686	0.760	0.074	0.760	0.819	0.066
T=1/3	0.889	0.906	0.020	0.833	0.874	0.019	0.938	<b>0.942</b>	0.016	0.696	0.769	0.072	0.758	0.825	0.064
T=1/4	0.893	0.908	<b>0.018</b>	0.832	0.870	<b>0.018</b>	0.934	0.933	<b>0.016</b>	0.693	0.768	0.074	0.756	0.822	<b>0.063</b>
T=1/5	<b>0.894</b>	0.906	0.020	<b>0.836</b>	<b>0.880</b>	0.019	0.935	0.940	0.017	<b>0.699</b>	<b>0.774</b>	<b>0.071</b>	<b>0.767</b>	<b>0.831</b>	0.066
T=1/10	0.888	0.906	0.019	0.836	0.876	0.018	<b>0.940</b>	0.939	0.016	0.698	0.769	0.071	0.738	0.812	0.070

TABLE II  
PROOFS REGARDING THE EFFECTIVENESS OF OUR MOTION QUALITY PERCEPTION MODULE (MQPM). THE QUANTITATIVE METRICS INCLUDE THE MAXF (LARGER IS BETTER), MEANF (LARGER IS BETTER), ADPF (LARGER IS BETTER), S-MEASURE (LARGER IS BETTER) AND MAE (SMALLER IS BETTER). BY USING THE MQPM AS THE INDICATOR, THOSE FRAMES WHICH ARE PREDICTED TO HAVE HIGH-QUALITY MOTIONS CAN OUTPERFORM OTHER FRAMES SIGNIFICANTLY, IN WHICH WE CHOOSE THE S<sub>SSAV</sub> [9] AS THE TARGET SOTA METHOD FOR EXAMPLE HERE.

Quality	Frames with High-quality Motions (HQ)					Frames with Low-quality Motions (LQ)				
Metric	maxF	meanF	adpF	S-M	MAE	maxF	meanF	adpF	S-M	MAE
Davis [39]	0.884	0.840	0.800	0.906	0.022	0.828	0.782	0.719	0.875	0.034
Segv2 [48]	0.864	0.808	0.834	0.881	0.024	0.780	0.726	0.709	0.852	0.024
DAVSOD [9]	0.653	0.621	0.626	0.753	0.080	0.642	0.611	0.611	0.738	0.086
Visal [49]	0.883	0.850	0.832	0.910	0.025	0.938	0.895	0.841	0.945	0.014
VOS [50]	0.767	0.739	0.749	0.815	0.073	0.734	0.697	0.700	0.816	0.074
Total	0.810	0.772	0.768	0.853	0.045	0.784	0.742	0.716	0.845	0.046

TABLE III  
COMPONENT QUANTITATIVE EVALUATION RESULTS. THE QUANTITATIVE METRICS INCLUDE THE MAXF (LARGER IS BETTER), S-MEASURE (LARGER IS BETTER) AND MAE (SMALLER IS BETTER), MORE DETAILS CAN BE FOUND IN SEC. IV-D.

Dataset	Davis			Segv2			Visal			DAVSOD			VOS		
Metric	maxF	S-M	MAE	maxF	S-M	MAE	maxF	S-M	MAE	maxF	S-M	MAE	maxF	S-M	MAE
MS Baseline	0.798	0.854	0.044	0.648	0.760	0.054	0.627	0.738	0.079	0.450	0.613	0.148	0.405	0.566	0.167
MS+MQPM	0.784	0.844	0.043	0.656	0.761	0.053	0.688	0.774	0.075	0.488	0.632	0.143	0.501	0.617	0.161
MS+MQPM	0.814	0.866	0.032	0.760	0.832	0.028	0.745	0.809	0.051	0.569	0.685	0.107	0.627	0.702	0.108
MS+MQPM+SOTA	<b>0.894</b>	<b>0.906</b>	<b>0.020</b>	<b>0.836</b>	<b>0.880</b>	<b>0.019</b>	<b>0.935</b>	<b>0.940</b>	<b>0.017</b>	<b>0.699</b>	<b>0.774</b>	<b>0.071</b>	<b>0.767</b>	<b>0.831</b>	<b>0.066</b>

training set contains a large number of such cases.

Meanwhile, we have noticed that there exists a large number of consecutive frames in the testing VSOD set (almost 30%) which are tend to be predicted as the ones containing high-quality motions. Since these consecutive frames usually share similar spatial appearance in general, it will easily lead to an over-fitted appearance model if we use all these frames during the up-coming training.

So, due to the above mentioned issues, we propose a novel filtering scheme, aiming to exclude the less-trustworthy or redundant training instances, see below.

**1)** For each frame in the new training set, we measure the consistency degree (we choose the S-Measure, but not limited to it) between its motion saliency map and the VSOD result produced by the target SOTA method.

**2)** For each T frames in the new training set, only one frame with the largest consistency degree—this consistency degree is usually positively correlated to the trustworthy degree regarding the VSOD predictions made by the target SOTA method, will be remained (see the detailed ablation study on

T in Table I).

**3) New Round Of Network Training:** Once the new training set has been constructed, we will conduct a new round of network training on it. However, we can not directly retrain the target SOTA model using this new training set, because it only consists of individual video frames without any temporal information; i.e., our new training set only preserves spatial information, while the SOTA models need to be fed by both spatial and temporal information. So, we choose to set up a completely new model with an identical network structure to the localization branch demonstrated in Fig. 4, and this new model will be trained over this new training set by using the common thread supervised training protocol (Eq. 6), and its output will be our final VSOD results with much improved performance compared with the target SOTA method.

Specifically, though this new round of training requires additional time cost, the performance gain can still benefit scenarios without speed requirements.

TABLE IV  
QUANTITATIVE COMPARISONS WITH CURRENT SOTA METHODS. THE TOP THREE RESULTS ARE MARKED BY RED, GREEN AND BLUE, RESPECTIVELY.

Dataset	Metric	Ours	2020		2019				2018			2017		
			PCSA [13]	LSTI [47]	SSAV [9]	MGA [12]	COS [46]	CPD [8]	PDBM [33]	MBNM [51]	SCOM [52]	SFLR [53]	SGSP [54]	STBP [52]
Davis [39]	maxF	<b>0.894</b>	<b>0.880</b>	0.850	0.861	<b>0.892</b>	0.875	0.778	0.855	0.861	0.783	0.727	0.655	0.544
	S-M	<b>0.906</b>	<b>0.902</b>	0.876	0.893	<b>0.910</b>	<b>0.902</b>	0.859	0.882	0.887	0.832	0.790	0.692	0.677
	MAE	<b>0.020</b>	<b>0.022</b>	0.034	<b>0.023</b>	<b>0.023</b>	<b>0.020</b>	0.032	0.028	0.031	0.064	0.056	0.138	0.096
SegV2 [48]	maxF	<b>0.836</b>	0.810	<b>0.858</b>	0.801	<b>0.821</b>	0.801	0.778	0.800	0.716	0.764	0.745	0.673	0.640
	S-M	<b>0.880</b>	<b>0.865</b>	<b>0.870</b>	0.851	<b>0.865</b>	0.850	0.841	0.864	0.809	0.815	0.804	0.681	0.735
	MAE	<b>0.019</b>	0.025	0.025	<b>0.023</b>	0.030	<b>0.020</b>	<b>0.023</b>	0.024	0.026	0.030	0.037	0.124	0.061
Visal [49]	maxF	0.935	<b>0.940</b>	0.905	0.939	0.933	<b>0.966</b>	<b>0.941</b>	0.888	0.883	0.831	0.779	0.677	0.622
	S-M	0.940	<b>0.946</b>	0.916	<b>0.943</b>	0.936	<b>0.965</b>	0.942	0.907	0.898	0.762	0.814	0.706	0.629
	MAE	<b>0.017</b>	<b>0.017</b>	0.033	0.020	<b>0.017</b>	<b>0.011</b>	<b>0.016</b>	0.032	0.020	0.122	0.062	0.165	0.163
DAVSOD [9]	maxF	<b>0.699</b>	<b>0.655</b>	0.585	0.603	<b>0.640</b>	0.614	0.608	0.572	0.520	0.464	0.478	0.426	0.410
	S-M	<b>0.774</b>	<b>0.741</b>	0.695	0.724	<b>0.738</b>	0.725	0.724	0.698	0.637	0.599	0.624	0.577	0.568
	MAE	<b>0.071</b>	<b>0.086</b>	0.106	0.092	<b>0.084</b>	0.096	0.092	0.116	0.159	0.220	0.132	0.207	0.160
VOS [50]	maxF	<b>0.767</b>	<b>0.747</b>	0.649	<b>0.742</b>	0.735	0.724	0.735	0.742	0.670	0.690	0.546	0.426	0.526
	S-M	<b>0.831</b>	<b>0.827</b>	0.695	<b>0.819</b>	0.792	0.798	0.818	0.818	0.742	0.712	0.624	0.557	0.576
	MAE	<b>0.066</b>	<b>0.065</b>	0.115	0.073	0.075	<b>0.065</b>	<b>0.068</b>	0.078	0.099	0.162	0.145	0.236	0.163

## IV. EXPERIMENTS

### A. Datasets

We have evaluated our method on five widely used public available datasets, including Davis [39], Segtrack-v2 [48], Visal [49], DAVSOD [9], and VOS [50].

- Davis dataset contains 50 video sequences with 3455 frames in total, and most of its sequences only contain moderate motions.
- Segtrack-v2 dataset contains 13 video sequences (exclude the penguin sequence) with 1024 frames in total, containing complex backgrounds and variable motion patterns, which is more challenging than the Davis dataset generally.
- Visal dataset contains 17 video sequences with 963 frames in total, and this dataset is a relatively simple one than others.
- DAVSOD dataset contains 226 video sequences with 23938 frames in total, which is the most challenging dataset in the field, involving various object instances, different motion patterns, and saliency shifting between different objects.
- VOS dataset contains 40 video sequences with 24177 frames in total, yet only 1540 frames were annotated well, in which the sequences are all obtained in indoor scenes.

### B. Implementation Details

We have implemented our method on a PC with an Intel(R) Xeon(R) CPU, Nvidia GTX2080Ti GPU (with 11G RAM) and 64G RAM. We use the DAVIS-TR [39] as the initial training set to train our motion quality perception model (MQPM). Also, an ADAM optimizer [55] is applied to update the network parameters. We set the batch size to 8 which takes almost all GPU memory. The initial learning rate is set to 10e-3. To avoid over-fitting problem, we have adopted the random horizontal flips for data augmentation.

### C. Evaluation Metrics

In order to accurately measure the consistency between the predicted VSOD and the manually annotated ground truth, we adopt three common used evaluation metrics, including the maximum F-measure value (maxF) [42], the mean absolute error (MAE) [41], and the structure measure value (S-measure) [43].

### D. Component Evaluation

We have conducted an extensive component evaluation to verify the effectiveness of our proposed motion quality perception module (MQPM), and the quantitative results can be found in Table III. Meanwhile, the corresponding qualitative demonstrations regarding this component evaluation can be found in Fig. 7.

As is shown in Table III, the performance of the learned motion saliency, which is denoted by “MS” and it can be obtained via  $\Theta(\text{OF}_i)$  as mentioned in Eq. 1, have exhibited the worst performance in all the adopted metrics. Then, by using our MQPM (Sec. III-B) to formulate a new training set (MS will be applied as the pseudo-GTs), the overall performance can be improved significantly (denoted by “MS+MQPM”), e.g., the maxF metric value in the VOS dataset has been increased from 40.5% to 62.7%. Notice that we can not achieve such performance improvements via the randomly assembled key frames from the training set, and we denote such implementation as “MS-MQPM”, of which the overall performance is quite similar to the original MS baseline. For example, in the *breakdance* video sequence of the Davis testing set, the MQPM has selected 16 high-quality key frames. For fair comparison, the “MS-MQPM” randomly select 16 frames as the key frames.

Since the object boundaries are usually blur in the MS baseline, the overall performance of the above re-trained model (i.e., “MS+MQPM”) is limited. Thus, we further resort our data filtering strategy (Sec. III-C2) to introduce the target SOTA results as the high-quality pseudo-GTs, of which the corresponding results are shown in the last row of Table III



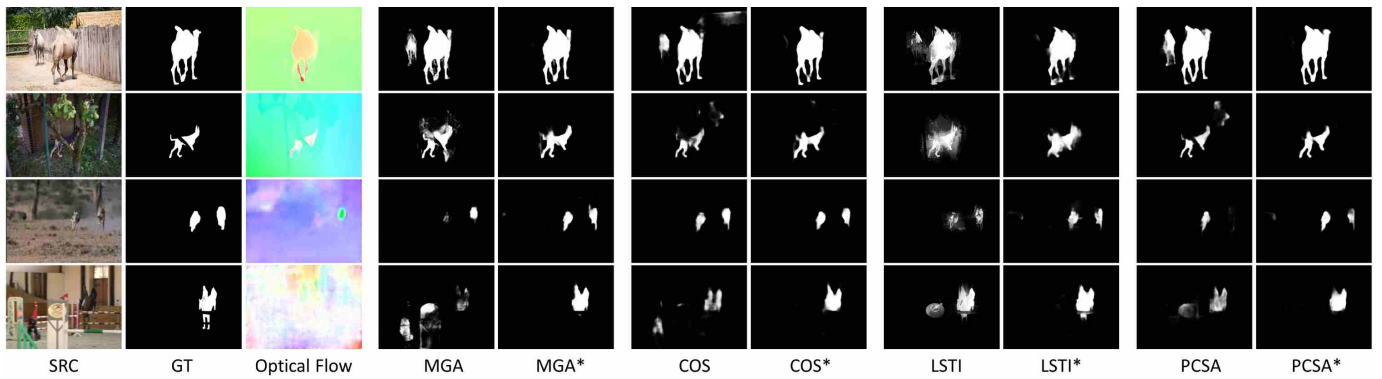


Fig. 6. Qualitative comparisons between several most representative target SOTA methods and the corresponding VSOD results after using our novel learning scheme.

TABLE V  
QUANTITATIVE COMPARISONS OF SEVERAL MOST REPRESENTATIVE SOTA METHODS (SSAV19, MGA19, COS19, LSTI20, AND PCSA20) VS. THEIR IMPROVED RESULTS BY USING OUR NOVEL LEARNING SCHEME.

Dataset	Metric	SSAV[9]	SSAV*	MGA[12]	MGA*	COS[46]	COS*	LSTI[47]	LSTI*	PCSA[13]	PCSA*
Davis [39]	maxF	0.861	<b>0.894</b>	<b>0.892</b>	<b>0.900</b>	0.875	<b>0.892</b>	0.850	0.863	0.880	<b>0.894</b>
	S-M	0.893	0.906	<b>0.910</b>	<b>0.914</b>	0.902	0.909	0.876	0.889	0.902	<b>0.909</b>
	MAE	0.023	0.020	0.023	<b>0.018</b>	0.020	<b>0.017</b>	0.034	0.024	0.022	<b>0.019</b>
SegV2[48]	maxF	0.801	<b>0.836</b>	0.821	0.835	0.801	0.815	<b>0.858</b>	<b>0.862</b>	0.810	0.835
	S-M	0.851	<b>0.880</b>	0.865	<b>0.882</b>	0.850	0.866	0.870	<b>0.891</b>	0.865	<b>0.880</b>
	MAE	0.023	<b>0.019</b>	0.030	0.028	0.020	<b>0.018</b>	0.025	<b>0.016</b>	0.025	0.020
Visal[49]	maxF	0.939	0.935	0.933	0.933	<b>0.966</b>	<b>0.956</b>	0.905	0.916	0.940	<b>0.942</b>
	S-M	0.943	0.940	0.936	0.931	<b>0.965</b>	<b>0.955</b>	0.916	0.928	<b>0.946</b>	<b>0.946</b>
	MAE	0.020	0.017	0.017	0.015	<b>0.011</b>	<b>0.010</b>	0.033	0.022	0.017	<b>0.014</b>
DAVSOD[9]	maxF	0.603	<b>0.699</b>	0.640	<b>0.672</b>	0.614	0.643	0.585	0.627	0.655	<b>0.680</b>
	S-M	0.724	<b>0.774</b>	0.738	<b>0.755</b>	0.725	0.736	0.695	0.718	0.741	<b>0.751</b>
	MAE	0.092	<b>0.071</b>	0.084	<b>0.075</b>	0.096	0.086	0.106	0.093	0.086	<b>0.077</b>
VOS[50]	maxF	0.742	<b>0.767</b>	0.735	<b>0.755</b>	0.724	<b>0.758</b>	0.649	0.690	0.747	0.758
	S-M	0.819	<b>0.831</b>	0.792	0.811	0.798	0.810	0.695	0.722	<b>0.827</b>	<b>0.824</b>
	MAE	0.073	0.066	0.075	0.066	<b>0.065</b>	<b>0.063</b>	0.115	0.101	<b>0.065</b>	<b>0.057</b>

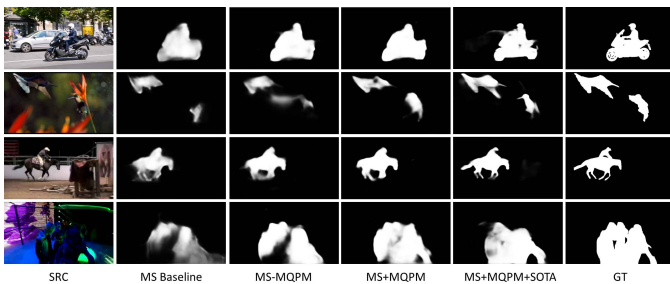


Fig. 7. The corresponding qualitative demonstrations regarding the component evaluations in Table III, in which the “MS+MQPM+SOTA” has achieved the best performance.

with the highest scores in all metrics, showing the effectiveness of our data filtering strategy.

Also, it should be noted that we have simply chosen the SSAV [9] as the target SOTA method here, because the off-the-shelf SSAV model was pre-trained using the identical training set as our method, which can avoid the data leakage problem.

### E. Ablation Study

As we have mentioned in Sec. III-C2, there are almost 30% video frames in the original testing set which will be predicted to contain high-quality motions (we abbreviate it as the high-quality frames). Due to the reasons we have mentioned in

Sec. III-C2, we believe that it is time-consuming and not necessary to use all these high-quality frames to start a new round of training. Thus, the main purpose of our data filtering strategy is to automatically keep a small subgroup of high-quality frames as the final training set.

Thus far, we have conducted an extensive ablation study regarding the parameter  $T$ , and the detailed results can be found in Table I. We choose  $T = \{1/10, 1/5, 1/4, 1/3, 1/2, 1\}$  respectively, in which  $T = 1$  means to use all those high-quality video frames as the new training set, and  $T = 1/5$  denotes only one frame with the largest consistency degree will be remained for each 5 consecutive high-quality frames. As is shown in Table I, the overall performance of our method is moderately sensitive to the choice of  $T$ , in which the overall performance via  $T = 1/5$  have exhibited the best performance in general, and a clear performance degradation can be found when we assign  $T = 1/10$ . So, we set  $T = 1/5$  as the optimal choice to strike the trade-off between performance and efficiency.

### F. Comparisons to the SOTA methods

We have compared our method with 12 most representative SOTA methods, including PCSA20 [13], LSTI20 [47], SSAV19 [9], MGA19 [12], COS19 [46], CPD19 [8],

TABLE VI

RUNTIME COMPARISONS, WHERE WE HAVE EXCLUDED THE TRAINING TIME (I.E., THE FPS PROVIDED HERE IS ONLY THE INFERENCE SPEED), BECAUSE THE TRAINING PROCEDURE MAY ONLY NEED TO BE CONDUCTED ONLY ONCE FOR MANY VIDEO SALIENCY BASED SUBSEQUENT APPLICATIONS. ALSO, OUR METHOD TAKES ABOUT 80S TO CONSTRUCT THE NEW TRAINING SET, AND ANOTHER 600S TO CONDUCT THE FINE-TUNING IN 5 EPOCHES (THIS WILL VARY WITH THE TRAINING SET SIZE); FOR A SINGLE TESTING FRAME, IT TAKES ABOUT 0.03S TO INFERENCE SOD RESULT.

Methods	Ours	PCSA20 [13]	LSTI20 [47]	SSAV19 [9]	MGA19 [12]	COS19 [46]	PDBM18 [33]	SCOM18 [56]	SFLR17 [53]	SGSP17 [54]
FPS	33	<b>110</b>	0.7	20.0	14.0	0.4	20.0	0.03	0.3	0.1
Platform	GTX2080Ti	GTXTitanXp	GTX1080Ti	GTXTitanX	GTX2080Ti	GTX2080Ti	GTXTitanX	GTXTitanX	GTX970	CPU

PDBM18 [33], MBNM18 [51], SFLR17 [53], SGSP17 [54], STBP17 [52] and SCOM18 [56].

As is shown in Table IV, all quantitative results have indicated that our method (we take the SSAV as the target SOTA model here) have significantly outperformed these compared SOTA methods for all tested datasets excepting the Visal dataset, showing the performance superiority of our method. In fact, the Visal dataset may be a bit different to other datasets, i.e., the Visal dataset is dominated by color information, in which the motion clues are usually at the second place to determine the true saliency. As a result, the COS19, which is heavily rely on the spatial domain, has exhibited the best performance in the Visal dataset. Also, we have provided the qualitative comparisons in Fig. 5, where our VSOD results are more consistent to the GT than those compared SOTA methods.

Moreover, our method can be applied to any other SOTA VSOD methods to get its performance further improved. To show such advantage, we have provided the direct comparisons between several most representative SOTA methods and their improved versions after using our learning scheme. As is shown in Table V, our method can make averagely 5% performance improvement generally and almost 9.6% regarding the best case (maxF), and the corresponding qualitative comparisons can be found in Fig. 6.

Also, we have conducted the running time comparisons to the SOTA methods in Table VI, in which our method has achieved the real-time speed with 33 FPS during the inference phase. Although our total time is a bit time-consuming, there are still advantages compared to other methods.

## V. CONCLUSION

In this paper, we have proposed a universal scheme to boost the SOTA methods within a semi-supervised manner. The key components in our method include: **1)** The motion quality perception module, which was used to select a subgroup of high-quality frames from the original testing set to construct a new training set; **2)** Data filtering scheme, which was used as a double-check to ensure the overall quality of the newly constructed training set. We have conducted an extensive quantitative evaluation to respectively show the effectiveness regarding these two components.

## REFERENCES

- [1] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, and F. Xu, "3d room layout estimation from a single rgb image," *IEEE Trans. Multimedia (TMM)*, 2020.
- [2] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2020.
- [3] K. Belloulata, B. Amina, and S. Zhu, "Object-based stereo video compression using fractals and shape-adaptive dct," *AEU-Int. J. Electron. Commun.*, vol. 68, pp. 687–697, 2014.
- [4] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognit. (PR)*, vol. 48, pp. 2885–2905, 2015.
- [5] Q. Fan, W. Luo, Y. Xia, G. Li, and D. He, "Metrics and methods of video quality assessment: A brief review," *Multimed. Tools Appl. (MTA)*, vol. 78, no. 22, pp. 31 019–31 033, 2019.
- [6] C. Chen, S. Li, H. Qin, and A. Hao, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognit. (PR)*, vol. 52, pp. 410–432, 2016.
- [7] C. Peng, Y. Chen, Z. Kang, C. Chen, and Q. Cheng, "Robust principal component analysis: A factorization-based approach with linear complexity," *Inf. Sci.*, vol. 513, pp. 581–599, 2020.
- [8] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3907–3916.
- [9] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8554–8564.
- [10] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 802–810.
- [11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4489–4497.
- [12] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 7274–7283.
- [13] Y. Gu, L. Wang, Z. Wang, Y. Liu, M. Cheng, and S. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020.
- [14] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag. (ISPM)*, vol. 35, no. 1, pp. 84–100, 2018.
- [15] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, vol. 28, no. 10, pp. 2473–2483, 2017.
- [16] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Trans. Image Process. (TIP)*, vol. 24, no. 8, pp. 2303–2316, 2015.
- [17] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 41, no. 4, pp. 815–828, 2019.
- [18] W. Wang, J. Shen, M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5968–5977.
- [19] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, 2019, pp. 199–204.
- [20] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6024–6033.
- [21] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3917–3926.
- [22] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 7479–7489.

- [23] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8779–8788.
- [24] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3395–3402.
- [25] F. Guo, W. Wang, J. Shen, L. Shao, J. Yang, D. Tao, and Y. Tang, "Video saliency detection using object proposals," *IEEE Trans. Cybern. (TCYB)*, vol. 48, no. 11, pp. 3159–3170, 2017.
- [26] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, "Bilevel feature learning for video saliency detection," *IEEE Trans. Multimedia. (TMM)*, vol. 20, no. 12, pp. 3324–3336, 2018.
- [27] F. Guo, W. Wang, Z. Shen, J. Shena, L. Shao, and D. Tao, "Motion-aware rapid video saliency detection," *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, 2019.
- [28] C. Chen, G. Wang, and C. Peng, "Structure-aware adaptive diffusion for video saliency detection," *IEEE Access.*, vol. 7, pp. 79 770–79 782, 2019.
- [29] R. Cong, J. Lei, H. Fu, M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, vol. 29, no. 10, pp. 2941–2959, 2018.
- [30] T. N. Le and A. Sugimoto, "Deeply supervised 3d recurrent fcn for salient object detection in videos," in *British Machine Vis. Conf. (BMVC)*, vol. 1, 2017, p. 3.
- [31] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3243–3252.
- [32] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process. (TIP)*, vol. 27, no. 1, pp. 38–49, 2017.
- [33] H. Song, W. Wang, S. Zhao, J. Shen, and K. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2018, pp. 715–731.
- [34] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "Accurate and robust video saliency detection via self-paced diffusion," *IEEE Trans. Multimedia. (TMM)*, vol. 22, no. 5, pp. 1153–1167, 2019.
- [35] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Trans. Multimedia. (TMM)*, vol. 22, no. 2, pp. 324–336, 2019.
- [36] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion," *IEEE Trans. Image Process. (TIP)*, vol. 29, pp. 4296–4307, 2020.
- [37] C. Chen, Y. Li, S. Li, H. Qin, and A. Hao, "A novel bottom-up saliency detection method for video with dynamic background," *IEEE Signal Processing Letters. (SPL)*, vol. 25, no. 2, pp. 154–158, 2017.
- [38] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4481–4490.
- [39] F. Perazzi, J. PontTuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 724–732.
- [40] D. Sun, X. Yang, M. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8934–8943.
- [41] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 733–740.
- [42] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 1597–1604.
- [43] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4548–4557.
- [44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2015, pp. 234–241.
- [46] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3623–3632.
- [47] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial-temporal information," *IEEE Trans. Image Process. (TIP)*, vol. 29, pp. 1090–1100, 2019.
- [48] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 2192–2199.
- [49] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process. (TIP)*, vol. 24, no. 11, pp. 4185–4196, 2015.
- [50] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process. (TIP)*, vol. 27, no. 1, pp. 349–364, 2017.
- [51] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C. Jay Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," in *Proc. IEEE Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 207–223.
- [52] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," *IEEE Trans. Image Process. (TIP)*, vol. 26, no. 7, pp. 3425–3436, 2016.
- [53] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process. (TIP)*, vol. 26, no. 7, pp. 3156–3170, 2017.
- [54] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, vol. 27, no. 12, pp. 2527–2542, 2016.
- [55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [56] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, "Scom: Spatiotemporal constrained optimization for salient object detection," *IEEE Trans. Image Process. (TIP)*, vol. 27, no. 7, pp. 3345–3357, 2018.