IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 10, 2019, accepted January 31, 2019, date of publication May 6, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899351

Structure-Aware Adaptive Diffusion for Video Saliency Detection

CHENGLIZHAO CHEN, GUOTAO WANG, AND CHONG PENG
College of Computer Science and Technology, Qingdao University, Qingdao 266071, China

Corresponding author: Chong Peng (pchong1991@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802215 and Grant 61806106 and in part by the Natural Science Foundation of Shandong Province under Grant ZR2019BF011 and Grant ZR2019QF009.

ABSTRACT This paper proposes a novel saliency model that reveals the long-term info to boost detection accuracy. The saliency estimation of conventional methods heavily depends on the locally revealed short-term info, and they could easily be trapped into imperfect configurations. In contrast, our method can take full consideration of common consistency of those *reliable* low-level predictions from the perspective of the entire video sequence. Meanwhile, we adopt a newly designed self-learning strategy which is guided by the low-rank analysis to adaptively reveal the long-term spatial–temporal video coherency. To avoid the error accumulations, we also propose a novel non-local descriptor to enhance the discriminative power of the feature space. Thus, the newly revealed the long-term info can be directly regarded as a trustful indicator to sustain additional low-rank analysis, which would serve as the basis toward selective fusion and significantly enhance the detection accuracy.

INDEX TERMS Video saliency detection, foreground modeling, spatial-temporal diffusion.

I. INTRODUCTION

The problem of video saliency detection aims to automatically detect the most salient object in the given video sequence. And such detections can be regarded as the ROI (region-of-interest) indicator to facilitate various downstream applications including person re-identification [1], video surveillance [2], and traffic control [3]. Different from the conventional image saliency which heavily rely on the spatial info as the exclusive saliency clue [4], [5], the incursion of video temporal info is the critical factor to obtain accurate video saliency. In fact, for downstream applications with saliency computation must be necessarily done online, e.g., the autopilot system [6], it is critical to accomplish the saliency aided environment perceiving immediately according to those newly arrived video streams. Thus, almost all the state-of-the-art methods [7]–[9] have to build their saliency model merely rely on the short-term info [10]–[12], which easily causing massive artifacts. However, for downstream applications with saliency accuracy at the highest priority, e.g., video summarization [13], the conventional in-available long-term info may become available, and this motivates us to utilize off-line manner to take full advantage of the long-term info for the accurate video saliency detection. To this end we summarize the current key technical challenges as follows.

First, the saliency exploring scope of the current state-of-the-art methods is too local to reveal long-term video saliency. In fact, rather than being limited in consecutive frames, the detection accuracy can be further boosted by considering more reliable long-term info as we mentioned before. Meanwhile, this revealed long-term info should also enable self-adaption to guide the current saliency prediction.

Second, simply applying the saliency corrections/amendments in the color spanned feature space easily causes the accumulation of false-alarm detections, especially when the salient foreground object shares similar color to its non-salient backgrounds. Thus, it is necessary to develop a much discriminative feature space which can enlarge the feature distance between the salient foreground object and its non-salient backgrounds.

Specially, our previous work FD17 [9] attempted to extend the saliency exploring extent in batch-wise manner via using low-rank guided fusion and diffusion procedure in color spanned feature space. However, because of the absence of the long-term info, FD17 encountered distinctive performance degradation when the majority of the intra batch low-level saliency are incorrect (see demonstrations in Fig. 1), not to mention the side-effects (e.g., the accumulation of

The associate editor coordinating the review of this manuscript and approving it for publication was Xian Sun.

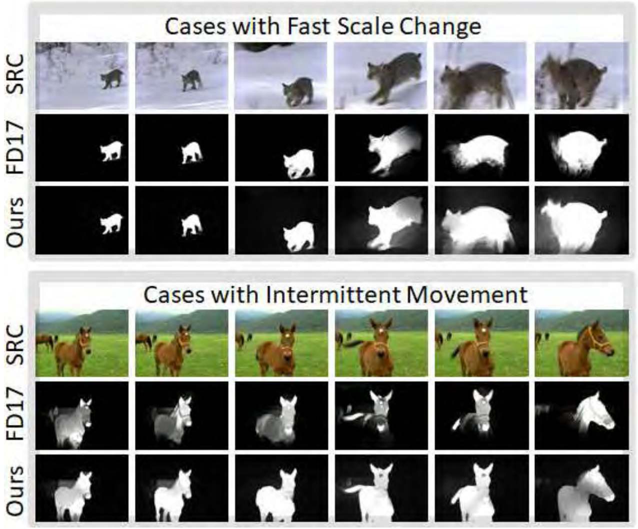

Cases with Fast Scale Change
SRC
FD17
Ours
Cases with Intermittent Movement
SRC
FD17
Ours

FIGURE 1. Motivation demonstration. Compared to FD17 [9], our method can well handle those fast scale change induced false negative detections and the intermittent movements induced hollow effects.

errors) brought by the low discriminatingly of the solely color-spanned feature space.

To ameliorate, our current research endeavors are aiming at iteratively establishing a much global self-learning based video saliency modeling solution to prevent the accumulation of the false-alarm detections while respecting the long-term info. To achieve this, we propose to introduce the newly designed structure-aware descriptor to span feature space with high discriminative power, which can well respects the spatial info while conserving the generalization power in temporal scale. Meanwhile, in order to correctly explore the long-term info, we propose to adaptively learn an additional appearance model from the low-level saliency predictions. Also, we utilize a series low-rank solutions to enable the adaptability of our learned appearance model, which is important for the saliency corrections/amendments to stay smoothness in temporal scale. Specifically, the salient contributions of this paper towards novel computational strategies can be summarized as follows:

- We design a novel descriptor to enhance the discriminative power of the color spanned sub feature space, which enable our subsequent saliency diffusion procedure to correct those previously ill-detections while avoiding the accumulation of errors.
- We propose a self-learning to reveal the long-term info to iteratively correct/amend those poor low-level saliency predictions, which suppose to exhibit much robust video saliency detections when the current low-level saliency is undergoing long period untrustful status.
- We integrate the low-rank analysis guided selective fusion strategy into our self-learning framework aiming to further constraint the temporal consistence of the detected video saliency.

II. RELATED WORK

A. IMAGE SALIENCY DETECTION METHODS

The central idea of image saliency detection is to extract the most eye-catching object that is significantly distinctive (i.e., uniqueness) from its non-salient surroundings. To represent the uniqueness, most of the earlier saliency methods directly employ global contrast as the saliency criterion, either in the raw color feature space [14] or in the frequency domain [15]. Following the same rationality, the improved multi-scale solutions dominate the saliency detection field for a long period of time, which explore the global saliency over the color information spanned feature spaces, including sparse dictionary based method [16], multi-level super-pixel feature based method [17], image boundary based method [18], etc. Although global contrast based saliency methods have achieved remarkable accuracy, they may easily miss some important sub-parts in the salient object due to the feature (color) overlapping between foreground and background. Then, local contrast methods are resorted to conquer this limitation, however, it tends to bring in hollow effect problems, which leave the inside regions of the salient object being undetected but assign high saliency value around the salient object boundaries [19]. The hybrid solutions (considering both local and global contrast) are also proposed to alleviate these limitations [20] by adopting more meaningful feature space [21], [22] and more discriminative descriptor [5], [23] to perform multi-scale contrast computation [24], [25]. Also, some high-level shape/structure based constraints [26] and priors [27] are introduced to sharpen the boundary of the salient object. Although the state-of-the-art image saliency detection methods have already achieved great success, they are still struggling to make trade-off between the local contrast and the global contrast. In particular, their detection performance over videos is extremely poor, and visual proofs (Fig. 10) can be found in our quantitative comparison in our experimental section.

B. FUSION-BASED VIDEO SALIENCY METHODS

Compared with the image saliency detection over spatial domain only, the incursion of video temporal/motion information is a critical factor which makes the video saliency challenging. The fusion based video saliency methods aim to integrate multiple saliency sources, which can be mainly classified into two categories such as color saliency clues and motion saliency clues, to achieve robust video saliency detections. Mahadevan and Vasconcelos [28] proposed to integrate spatial contrast computation into temporal scale, and then the obtained temporal-spatial contrast (i.e., center surrounded spatial-temporal region, followed and improved by [25]) was directly regarded as the spatial-temporal video saliency clue. Similarly, Seo and Milanfar [29] proposed to compute the contrast based saliency in pre-defined spatial-temporal surroundings directly. Although these methods [28], [29] considered both the spatial and temporal saliency clues simultaneously, neither of these clues is accurate enough to produce high quality video saliency detections due to the absence of the long-term spatial-temporal coherency. To overcome this limitation [30] resorted the Conditional Random Field (CRF) to fuse more saliency sources, e.g., the illumination contrast, achieving better video saliency detections than [28]. Also following the same principle, i.e., "the more, the better", [13] further explores the motion saliency from the temporal perspective, including the foreground object related velocity contrast and acceleration contrast, and the final video saliency is computed via the "additive" fusion [31] in a multi-scale manner, which on average combines multiple saliency sources to boost the detection accuracy. In fact, the above methods still rely on two basic video saliency sources (i.e., the color saliency and motion saliency), yet with different formulations or combinations, which obviously are encountering unsolvable performance bottleneck due to their bootstrap motivation. Thus, rather than roughly combining multiple basic saliency sources for the final video saliency detection, the most recent fusion based works are mainly focusing on the designation of the selective strategy, which estimates the video saliency by automatically seeking an appropriate balance respectively from the above-mentioned basic saliency clues. For example, Fang *et al.* [12] adopted the entropy as the major indicator to evaluate the performance of individual basic saliency clues, and then the final video saliency were designed to strictly bias toward those saliency clues with high entropy measure. Most recently, Liu *et al.* [32] regarded the mutual consistence as the criterion to guide the collaborative interaction and selection between temporal and spatial spaces, which achieved remarkable performance improvement. Although these newly-designed selection based fusion strategies [12], [32] can produce robust saliency detection results, unfortunately, the continuity nature of the movement (i.e., the consistence of the spatial-temporal information between consecutive video frames, named as the spatial-temporal coherency) is totally neglected, and massive false-alarm detections frequently occur when both spatial and temporal saliency clues are undergoing unreliable condition during a long perior of time.

C. VIDEO SALIENCY METHODS GUIDED BY SPATIAL-TEMPORAL COHERENCY

Almost all the spatial-temporal coherency guided video saliency methods follow the rationality that, the movement trajectories of the salient foreground object are frequently characterized by spatial-temporal smoothness. Thus, long-term modeling with appropriate update is the most intuitive solution to explore the spatial-temporal coherency in consecutive video frames. From the perspective of scene modeling, background subtraction based salient motion/change detection methods [2], [33], [34] have been well studied in recent years, whose central idea is to utilize low-rank decomposition [35] to automatically separate the salient foreground (i.e., the sparsity component) from the non-salient background (i.e., the low-rank component) by seeking spatial-temporal coherency, and the sparsity measure is regarded as the unique indicator to locate the motions/changes. Although plausible detection performance has been observed for the stationary videos, these modeling based methods frequently become incapable for non-stationary videos [34] due to the absence of pixel-wise correspondence in consecutive video frames. To ameliorate, either frame-level affine registration [36] or background tracking strategy [37] is integrated into the low-rank revealing process to convert the non-stationary scenarios to relatively stationary ones, however, the obstinate challenges still exist when the input video sequences only have limited frames, because it can heavily impacts the robustness of the estimated background model and leads to poor performance. Different from the above-mentioned modeling methods, which mainly model the spatial-temporal coherency of the non-salient backgrounds, [10] proposes to utilize the foreground spatial-temporal information to construct their attention model, which fully takes the advantages of the motion continuity to eliminate false-alarm detections. So the limitations of video saliency detection over non-stationary video have been significantly alleviated with magnificent performance improvement in stationary scenarios. Similarly, Li *et al.* [31] proposed to utilize the newly-designed kernel regression to explore the local spatial-temporal coherency, whose hidden rationale is to seek the common consistencies of the foreground object in short-term video in a batch-wise way. Kim *et al.* [8] regarded the graph model (to be built over short-term video frame batches) based stationary status as the video saliency clue, achieving plausible performance. Actually, the core rationality of the above batch-wise spatial-temporal coherency method is to constrain the detected video saliency of the local neighboring frames (i.e., consecutive frames in short-term frame batches) to stay spatial-temporal consistence. It follows the assumption that, the overall appearance of the foreground object should remain roughly identical. Thus, the obtained spatial-temporal coherency is localized in the predefined short-term frame batches, and their performance constantly struggles to deal with the trade-off between the predefined frame batch size and the intensity level of the varying foreground object. So, [38] proposes to utilize a newly-designed graph model (considering the unbound spatial-temporal coherency of the foreground object) to automatically conduct the video saliency detections, which was further followed and improved by Wang *et al.* [7], [39]. Although these graph model based video saliency methods have achieved remarkable performance improvement, the graph model solution easily causes the accumulation of false-alarm detections, because its un-bounded saliency expansion lacks of a mechanism to suppress the non-salient backgrounds while enhancing the salient foreground object. This paper will explore the global spatial-temporal coherency in an un-bounded manner to avoid the accumulation problem of false-alarm detections.

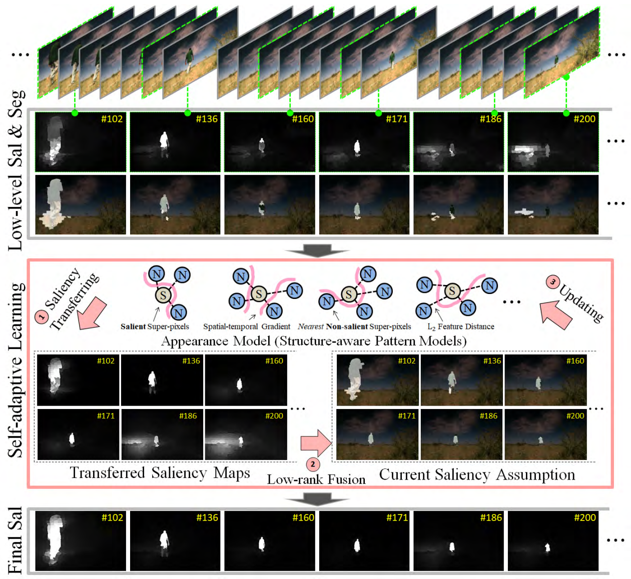


FIGURE 2. Overview of our proposed video saliency method.

III. METHOD OVERVIEW

As we can see in Fig. 2, our method mainly consists of three components: (1) structure-aware saliency transfer (**Mark 1**); (2) low-rank analysis guided selective fusion (**Mark 2**); (3) appearance model update (**Mark 3**).

Our method first decomposes the original input video sequence into video batches (identical to our previous work FD17 [9]), and then compute **Spatial-temporal Gradient** (Fig. 13) guided novel low-level saliency, i.e., the color saliency and the motion saliency, which will be discussed in **APPENDIX**. Based on the obtained low-level saliency clues, we propose to automatically formulate structure-aware **Pattern Models** (middle row in Fig. 2) to facilitate our self-adaptive learning iterations, which mainly include the following components:

A: automatically obtain the binary segmentations as the foreground object mask (**Algorithm 1** in Sec. IV-A) which can be regarded as the intermediate saliency assumption to facilitate the computation of our structure-aware descriptor (Sec. IV-B) in the current learning iteration;

B: detail our newly-proposed structure-aware descriptor to enhance the discriminative power of color-based feature subspace, which is extremely important for the following *Saliency Transfer Procedure* (see **Mark 1** in Fig. 2), and more details can be found in **Algorithm 2** in Sec. IV-B;

C: utilize our newly-designed saliency transfer (details can be found in Sec. V-C) to amend the previous detection results in an iterative fashion via *Low-rank Analysis Guided Selective Fusion* (see **Mark 2** in Fig. 2), which automatically performs the selective combination between the low-level predictions and the transferred saliency according to the sparsity measure of our low-rank analysis. It will be further discussed in Sec. V-D;

D: the learned pattern models and the foreground mask will be iteratively updated after each self-learning procedure, see **Mark 3** in Fig. 2 and more details can be found in Sec. V-E.

IV. STRUCTURE-AWARE DESCRIPTOR

Given an input video sequence, we follow our previous work (FD17 [9]) to perform the low-level saliency estimation in frame-wise manner, and the details can be found in **Appendix** of this manuscript.

A. ADAPTIVE BINARY LABELING

Since we have already obtained the low-level saliency (**LS**, Eq. A5) for each video frame in the given video sequence, it is indispensable to conduct binary segmentation in the current learning iteration as the preliminaries of our structure-aware description (details can be found in Sec. IV-B). However, due to the varying appearance of the salient foreground object, it is not advisable to adopt hard threshold (i.e, the most commonly used: 2 × $mean(LS)$) to perform uniform binary labeling. Although the appearance of the salient foreground may undergo exquisite variation extensively, the appearance-varying tendency frequently follows the spatial-temporal coherency, which is highly constrained in temporally neighbored short-term frame batches. Thus, we propose to employ the shape prior (**SP**) as the adaptive threshold to iteratively perform the coarse-to-fine binary labeling batch-wisely.

Here we demonstrate the detailed computation of the shape prior in Eq. 1, where the subscript j and k respectively represent the frame index and batch index, **STG** denotes the spatial-temporal gradient map, d_i denotes the location of the i-th non-zero element in **STG**, $C_{k,j,1}$ denotes the single center location of the 1st-round K-means clustering, $C_{k,j,2}$ and $C_{k,j,3}$ represent the center locations of the 2nd-round K-means clustering, $f(\cdot)$ denotes the distance filter, which only considers those distances between 5-th and 95-th percentiles of the observed l_2 distances distribution, and hard threshold α is empirically assigned to 0.25 × $min\{W, H\}$, where W and H separately denotes the width and height of the given video frame.

$$SP_j \leftarrow \begin{cases} \underbrace{f(\frac{\sum_i ||d_i - C_{k,j,1}||_2}{||STG_j||_0})}_{if\ \ |C_{k,j,1}-C_{k,j,2}|+|C_{k,j,1}-C_{k,j,3}|\leq\alpha} \\ \underbrace{f(\frac{\sum_i ||d_i - C_{k,j,2}||_2}{2 \times ||STG_j||_0}) + f(\frac{\sum_i ||d_i - C_{k,j,3}||_2}{2 \times ||STG_j||_0})}_{otherwise}, \end{cases} \tag{1}$$

Also the pseudo-code of our adaptive labeling procedure to obtain the foreground mask (**FM**$\in \{0, 1\}^{W\times H}$) can be found in **Alg.1**, where $pos(\cdot)$ represents the binary labeling function that automatically assigns 1 to those non-zero elements and 0 to the remaining ones, β is the learning factor that is empirically set to 0.7 to avoid fluctuation induced inter-batch artifacts. In fact, the underlying rationality of **Alg.1** is to constrain the binary labeling procedure iteratively to respect the estimated shape prior, yet the over-fitting problem could be

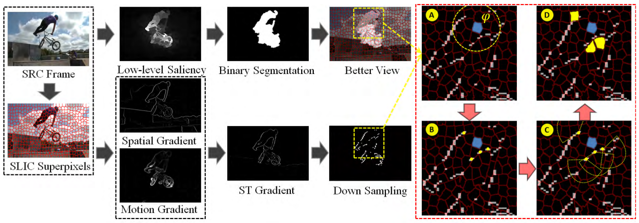


FIGURE 3. **Demonstration of our structure-aware pattern descriptor, where the given superpixel and its corresponding "non-salient" topology constraints are respectively marked with blue and yellow color in Structure-aware Description.**

```
Algorithm 1 Iterative Binary Labeling
    Input: Shape prior separately from $t-1$ batch
        and $t$ batch: $SP_{t-1}$, $SP_t$;
        Low-level saliency map of $i$ frame (Eq. A5): $LS_i$.
        Labeling threshold: $LT_{i-1} = 2 \times avg(LS_i)$;
        Total super-pixel number of $i-1$ frame: $n_i$;
    Output: Foreground mask of the i-th video frame: $FM_i$.
    Initialization:
        Hard threshold $T_i = LT_{i-1}$ at the beginning,
        or $\underbrace{T_i = 2 \times avg(FSal_i)}_{for\ the\ following\ iterations}$(Eq. 7).
    1. $SP_t = (1-\beta) \times SP_{t-1} + \beta \times SP_t$;
    While (1)
    2. $\underbrace{if((\pi \times SP_t^2)/(W \times H) - ||LS_i - T_i||_0/n_i < 0.1)}_{continue\ iteration;}$
    3. $T_i = T_i \times (1 - 0.05 \times sign((\pi \times SP_t^2)/(W \times H)$
           $-||LS_i - T_i||_0/n_i))$;
    End While
    4. $LT_i = (LT_{i-1} + T_i)/2$;
    5. $FM_i = pos(LS_i - LT_i)$;
```

avoided by adopting the slack parameter (i.e., 0.1 in step 2 of **Alg. 1**). Specifically, because the initial hard threshold LT is selected based on the low-level saliency, poor binary labeling result may be easily obtained at the early iterations, thus, the binary labeling result can be further improved in our following self-adaptive learning iterations.

B. NOVEL STRUCTURE-AWARE DESCRIPTOR

The technical foci of this paper are to utilize the self-adaptive learning process to capture the long-term common consistence of the salient foreground object. To achieve this, we propose to represent all the low-level saliency in the formation of explicit sub graph model, and we name it as Pattern Models (**PM**). Considering the robustness of the **STG** (Spatial-temporal Gradient map, see details in **Appendix**) in motion sensing, we propose to utilize the **STG** to adaptively locate its corresponding non-local regions. That is, for any binary labeling indicated salient super-pixel (blue regions in right part of Fig. 3), we attempt to locate multiple "non-salient" super-pixels (yellow regions in right part of Fig. 3) as a weak structure-aware constraint, and we empirically set its number to 3 according to the quantitative results toward different super-pixel size. We now summarize the detailed steps of our structure-aware descriptor in **Alg.2**.

```
Algorithm 2 Structure-Aware Descriptor
    Input: Spatial-temporal gradient map: $STG$;
        Super-pixel number: $n$;
        Labeling mask and SLIC super-pixel map: $FM$, $SM$;
        Total number of the salient super-pixels: $u$;
        The center location of the target super-pixel: $p_t$;
    Output: Pattern model of the t-th superpixel: $PM$.
    Initialization: $v = 2 \times u$.
    For $t = 1 : n_i$
      if $(FM(p_t) == 1)$
    1. Locate $v$ non-zero elements $\underbrace{\xi = \{\xi_1, \xi_1, ..., \xi_v\}}_{Mark\ B\ in\ Fig.\ 3}$
       in region $\varphi$($Mark\ A$ dash circle located at $p_t$);
    2. Select $v$ non-salient super-pixels $Q = \{q_1, q_2, ..., q_v\}$
       from the Sector Area, which anchors at each $\xi$
       with main direction $\vec{\zeta} = \xi - p_t$ ($Mark\ C$), while
       satisfying $FM(q_i) == 0$ and $arg\ \min_{q_i} ||q_i - \xi||_2$;
    3. Select the top $u$ super-pixel from Q
       according to $||Q - p_t||_2$;
    4. $PM \leftarrow [\ PM \quad \{SM(p_t) \ \ SM(Q)\}\ ]$ ($Mark\ D$);
      end if
    End For
```

In **Alg.2**, ξ (step 1) denotes the coordinates of non-zero elements in **STG**, and we also utilize the shape prior (Eq. 1) to control the radius of φ (i.e., 1.2 × SP), which is marked with yellow dash lines in sub-figure **A** of Fig. 3. Also, q (step 2) represent the average center location of the selected non-salient super-pixel indicated by foreground mask (FM, **Alg.1**), and we assign the radius of $\vec{\zeta}$ centered sector area ($\pm 35^o$) as $1.2 \times SP$, which is identical to the φ. Apparently, the obtained $PM \in \mathbb{R}^{(1+v) \times ColorDim+1}$ contains the local topology info coupled with its current saliency measure, wherein $v = 3$ denotes the number of selected topology constraints, $ColorDim = 5$ denotes the feature dimension of our adopted color info, including the *RGB* color info and the last two dimensions of *Lab* color info. The demonstration of the obtained pattern model can be found in sub-figure **D** of Fig. 3. Specifically, the newly formulated pattern models are informative enough to span feature space for our subsequent saliency transferring scheme, because both the *STG* and the foreground mask are mutually suppressed at the same time. Thus, whenever the foreground mask is incorrect, the discriminative power of the corresponding feature space would be degenerated if we solely consider the *STG*, and vice versa.

V. SELF-ADAPTIVE LEARNING

Since the purpose of our self-adaptive learning scheme is to utilize those good low-level predictions to amend the remaining bad ones, we propose to learn an *AM* (Appearance Model, Sec. V-B) to adaptively facilitate global video saliency

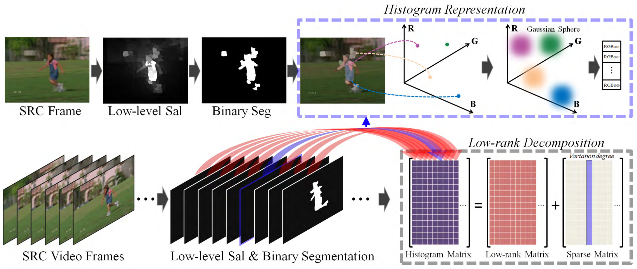


FIGURE 4. The pipeline of our low-rank analysis method.

prediction, which leverages multiple **PM** (Pattern Models, **Alg.2**) to capture the common consistence of the salient foregrounds in our structure-aware descriptor spanned feature space, i.e., $AM = \{PM_1, PM_2, ...\}$.

A. LONG-TERM LOW-RANK COHERENCY ANALYSIS

Observing the fact that the overall foreground appearance frequently stay unchanging within the limited video frames, we propose to further explore its common consistence by performing low-rank analysis to facilitate our previous mentioned low-level saliency amendment. To gain a better understanding, we demonstrate the overall pipeline of our low-rank analysis method in Fig. 4, wherein the current foreground mask indicated foreground regions are represented with the [10, 10, 10] histogram, which are dilated with 3-ring discrete Gaussian sphere ($variance = 10$) to ensure the generalization ability. And then, the foreground regions corresponded histograms are catenated into high-dimensional vector (m) to constitute the histogram matrix, i.e., the appearance matrix **D**. We now provide a brief introduction about the common thread low-rank analysis. Its main purpose is to decompose the input appearance matrix **D** into a low-rank part **L** and a sparse part **E**, as $D = L + E$. So the problem can now be formulated as:

$$\min_{L,E} \; rank(L) + \lambda ||E||_0 \quad s.t. \;\; D = L + E. \tag{2}$$

Since Eq. 2 is a non-convex optimization problem (which is NP-hard), however, it can be approximately solved via its relaxing convex envelope as:

$$\min_{L,E} \; ||L||_* + \lambda ||E||_1 \quad s.t. \;\; D = L + E. \tag{3}$$

Here $|| \cdot ||_*$ indicates the nuclear norm of **L**, and λ controls the sparsity measure, which is assigned to $\frac{1}{\sqrt{m}}$ following the suggestion of [34] and [35]. Of which, m represents the feature dimension of original data. Obviously, since the obtained low-rank component **L** represents the common consistence of the foreground object along the temporal axis, the column-wise l_1 norm of the revealed sparse component **E** is trustful enough to indicate those ill-detected low-level predictions.

B. APPEARANCE MODEL CONSTRUCTION

As aforementioned, the goal of our learning scheme is to establish an appearance model (**AM**) to record the common consistence of the detected salient foregrounds. However, the appearance model should be well initialized in advance due to the iterative nature of our learning procedure. Since the sparse component **E** (Eq. 3) is relatively trustful toward the assessment of low-level saliency predictions, we propose to batch-wisely locate one key frame with lowest sparsity degree ($||C_i(|E|)||_1$) as the anchor frame (Fig. 2, Step 4) and use its embraced pattern models to initialize the appearance model. After that, the appearance model will be automatically becoming self-contained, where each learning iteration starts with the saliency transferring (Sec. V-C) simultaneously at all the current anchor frames and ends with the updating procedure (Sec. V-E).

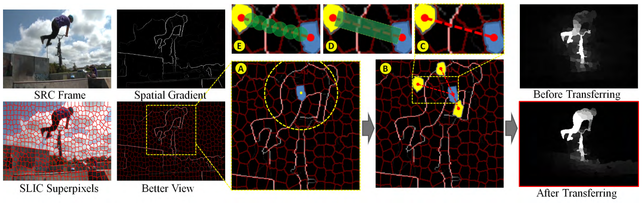


FIGURE 5. Demonstration of our structure-aware saliency transferring strategy.

C. SALIENCY TRANSFERRING

Actually, the key rationality of our saliency transferring procedure (Fig. 5) is to perform spatial-temporal saliency weighting according to the "similarity" degree between two neighbored **PM**s (Pattern Models). Here, we mainly consider the similarity between two spatial-temporal neighbored **PM**s from two aspects: the color info and the local topology info. Thus, the corresponding formulation of the above mentioned novel similarity measurements (ϖ) can be detailed as follows:

$$\begin{aligned} \varpi &= exp(-\omega_1 \cdot ||PM_1 - c_i||_2 \\ &\quad - \omega_2 \cdot \min_{j \in \psi}\{\sum_{k=2}^{v} exp(-G(\nabla_j)) \cdot ||PM_k - c_j||_2\}), \end{aligned} \tag{4}$$

where ω_1, ω_2 are the weighting parameters, the definition of v is identical to that in **Alg.2**, $c \in \mathbb{R}^{5\times1}$ denotes the color info (3RGB+2Lab), ∇ denotes the color gradient-like contour detections, and ψ (Eq. 5) shrinks the problem domain (i.e., radius of the searching range, see the yellow dash circle in Fig. 5**A**) to facilitate the minimum topology distance computation. Here parameter κ is identical to Eq. A2.

$$\begin{aligned} \psi &= \{||j - i||_2 \leq 1.1 \times \kappa\}, \\ &\qquad s.t. \; \theta \leq ||j - i||_2 \leq \max\{0.5 \times SP, \theta\}. \end{aligned} \tag{5}$$

Specifically, function $G(\cdot)$ in Eq. 4 returns the local maxima of the ∇ (spatial gradient) within the local neighborhood located along the two given superpixel. For example, the green rectangle region in Fig. 5**D** can be efficiently implemented via exhaustive search within multiple green circles (with uniform radius 15, Fig. 5**E**). In fact, the left part of Eq. 4 measuring the color similarity, and the right part respects

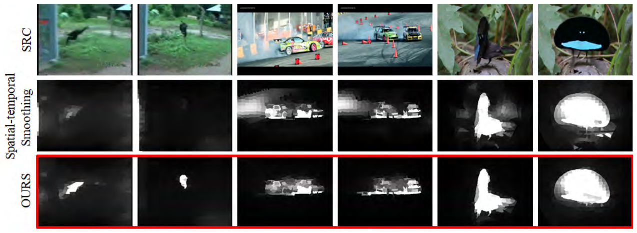


FIGURE 6. Demonstration of the advantages of our novel saliency transferring strategy.

to the consistence degree between two correlated topology info. And both the parameters ω_1 and ω_2 control the trade-off between these two components, which will be further discussed in Sec. VI-A. In practice, the optimal solution of the right part of Eq. 4 can be efficiently solved by Hungarian algorithm [40] in polynomial time. We propose to use CUDA acceleration to solve it in parallel by invoking an individual CUDA kernel for each target super-pixel.

So far the transferred saliency value ($TSal$) can be easily obtained via the following majority voting scheme (Eq. 6).

$$TSal_j = \frac{\sum_{i=1}^{M} \sigma_i \times \varpi_{i,j} \times Sal_i}{\sum_{i=1}^{M} \sigma_i \times \varpi_{i,j}}. \tag{6}$$

In Eq. 6, $\varpi_{i,j}$ denotes the computed similarity measure (using Eq. 4) between the i-th pattern model and the j-th pattern model, M denotes the total pattern model number in the appearance model, Sal_i denotes the saliency info provided by the i-th pattern model. And $\sigma_i = exp(-0.5 \times ||p_i, p_j||_2)$ is the spatial constraint, where p_i and p_j respectively represent the location of the corresponding pattern models.

The qualitative performance improvement brought by our saliency transferring strategy can be easily observed in Fig. 5 (please refer to the difference between the Before/After Transferring). We further demonstrate the advantages of our novel descriptor supported saliency transferring strategy in Fig. 6, wherein the accumulation of false-alarm detections can be easily observed from the method adopting spatial-temporal smoothing, while our method can handle this problem well. Also, the quantitative proofs can be found in the Sec. VI-A (Fig. 9b). After the saliency transferring procedure (Eq. 6), the newly-transferred saliency result will be applied to guide the low-level saliency amendment via our low-rank analysis guided fusion scheme (Sec. V-D).

D. LOW-RANK ANALYSIS GUIDED LOW-LEVEL SALIENCY AMENDMENT

In the current self-adaptive learning iteration, the quality of the $TSal$ (Transferred Saliency, Eq. 6) is strictly determined by one previously estimated prerequisites, i.e., the foreground mask. In fact, the estimated foreground mask tends to exhibit low accuracy at the earlier learning iterations, and this may directly result in performance degradation of our saliency transferring procedure. Therefore, for the current learning iteration, it is not advisable to directly regard the transferred saliency as the current low-level saliency assumption.

Being noticed the fact that the long-term low-rank coherency can well represents the trustful degree of the current low-level saliency, we propose to perform selective saliency fusion before we conduct new learning iteration. That is, we propose to fuse the current low-level saliency with the transferred saliency as the final saliency estimation in the current learning iteration. And our fusion procedure will bias toward the low-level saliency if the column-wise l_1 norm of the sparse component E (Eq. 3) exhibits small value. Therefore, we formulate our low-rank analysis guided saliency amendment as following:

$$\begin{aligned} \mathbf{FSal}_i &= N(\eta_1 \times \mathbf{TSal}_i + (1-\eta_1) \times \mathbf{LS}_i) \\ &\quad \odot ((1-\eta_2) \times N(\mathbf{TSal}_i) + \eta_2), \qquad (7) \\ \eta_1 &\leftarrow (1-\eta_2) \times N(||C_i(|\mathbf{E}|)||_1) + \eta_2. \qquad (8) \end{aligned}$$

Here the **FSal** denotes the fused video saliency result, η_2 (we empirically set it to 0.7) is the predefined parameter to suppress the influence of the element-wise Hadamard operation $\odot$, $N(\cdot)$ denotes the min-max [0, 1] normalization function. The latent rationality of the Hadamard operation in Eq. 7 is to further suppress those non-salient backgrounds while boosting the saliency measure of the common consistences. Hence, the parameter η_1 in Eq. 7 determines the bias tendency between the transferred saliency (**TSal**) and the low-level saliency (**LS**), which can be automatically assigned according to the sparsity measure by using Eq. 8. Of which, $C_i(\cdot)$ represents the column-wise selection operator, which returns the i-th column of the given input.

$$\eta_2 \leftarrow 0.7 \times \eta_2, \quad \textit{if} \;\; ||C_i(|E|)||_1 > \frac{2}{FN}\sum_{j=1}^{FN} ||C_j(|E|)||_1. \tag{9}$$

Specifically, for those video frames with extremely intensive scenario variations, i.e., a very large $||C_i(|E|)||_1$, it is reasonable to increase the bias degree toward the **TSal** as Eq. 9. FN in Eq. 9 denotes the total frame number.

E. APPEARANCE MODEL LEARNING AND UPDATING

In order to maintain the adaptability of the learned appearance model, we employ the residual between the **TSal** (Transferred Saliency, Eq. 6) and the previous estimated low-level saliency to jointly determine which pattern model should be updated. Also, inspired by sparse sensing theories [41], the appearance model can be sparsely represented by Gaussian-like random observations if the observation matrix satisfies the Johnson-Lindenstrauss lemma [42]. For example, following the sparsity lower-bound $m/log(m)$ (m represents the data dimension), we propose to normalize the above "residual" into Gaussian-like probability to control the updating procedure, and the performance improvement can be found in the bottom row of Fig. 7. Specifically, we utilize Eq. 10 to compute the original updating probability, wherein the similarity degree ϖ can be computed by Eq. 4, ψ and σ are

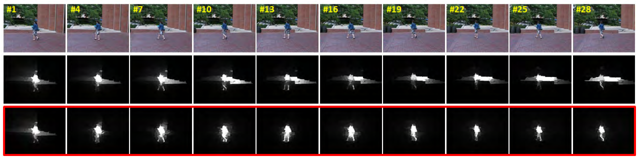


FIGURE 7. The middle row demonstrates the video saliency detection results without self-learning updating procedure, and the bottom row demonstrates the detection results with our novel self-learning updating procedure.

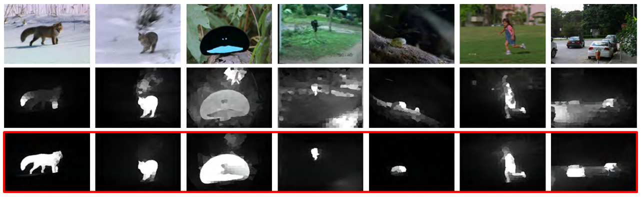


FIGURE 8. Demonstration of performance improvement brought by our novel learning scheme. The middle row demonstrates the low-level saliency results computed by Eq. A5, and the bottom row shows the video saliency performance of our newly-developed learning scheme.

identical to that in Eq. 5, and the **FSal** can be obtained via Eq. 7.

$$p_j = \frac{\sum\limits_{i\in\psi} |N(FSal_i^{t-1}) - N(TSal_i^t)| \times \sigma_i \times \varpi_i}{\sum\limits_{i\in\psi} \sigma_i \times \varpi_i}. \qquad (10)$$

To this end, all these computed probability (*prob*) in the appearance model can be represented as $prob = \{p_1, p_2, ..., p_{an}\}$, where *an* denotes the total pattern model number in the appearance model, and then, in order to transform *prob* into Gaussian-like formulation to meet the Johnson-Lindenstrauss lemma, we further normalize *prob* by using Eq. 11.

$$prob = \begin{cases} 0.5 \times N(prob_1) \\ \qquad if \;\; p_i < C_{\lfloor \frac{cn}{log(cn)} \rfloor}(Rank(N(prob_1))) \\ 0.5 \times N(prob_1) + 0.5, \quad otherwise, \end{cases} \qquad (11)$$

where $Rank(\cdot)$ denotes the forward ranking operation, $C_k(\cdot)$ returns top-k elements of the given input, $N(\cdot)$ is the normalization function, $\lfloor \cdot \rfloor$ represents the round down operation, and *cn* denotes the total pattern model number in the current video frame.

As discussed above, we propose to utilize the buffering strategy to discriminatively control the updating procedure to handle the motion related false-alarm detections. Thus, we introduce an additional parameter *buf* to buffer the updating procedure, so that the corresponding pattern model will be replaced by the currently computed novel pattern model if *buf* reaches zero, see details in Eq. 12.

$$buf_i = \begin{cases} buf_i - 1 & if \;\; prob_i > \tau \\ min(buf_i + 1, \delta) & otherwise, \end{cases} \qquad (12)$$

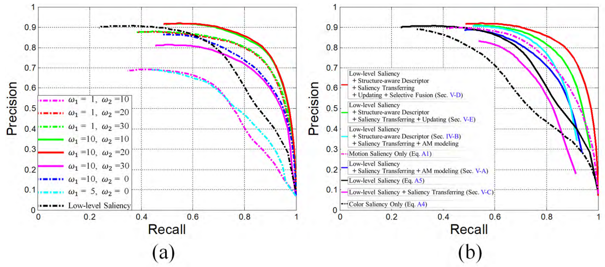


FIGURE 9. (a) PR (Precision-recall) curves of our method under different choices of ω_1 and ω_2 (Eq. 4); (b) PR curves of our method combining with different components, where the symbol (+) denotes the sub component combining.

where τ denotes a random value between [0, 1], $prob_i$ represents the i-th pattern model's updating probability (Eq. 11), and δ represents the initial buffer size, whose optimal choice will be discussed in Sec. VI-A. Besides, Fig. 8 demonstrates the performance improvement brought by our self-adaptive learning scheme.

VI. EXPERIMENTS AND EVALUATIONS

A. PARAMETER SETTINGS

In principle, there are three parameters having influence on the performance of our method: the trade-off parameters ω_1 and ω_2 (Eq. 4), the initial buffer size δ (Eq. 12) and the SLIC super-pixel size. As the first two parameters ($\omega_{1,2}$ and δ) can directly affect the performance of video saliency detection, in order to obtain the optimal overall performance, we comprehensively test their entire effects quantitatively (these two parameters are independently evaluated), and then determine the optimal selection of the detailed SLIC super-pixel size.

1) PARAMETER ω_1 AND ω_2

(Eq. 4). We quantitatively test the performance of these parameters to obtain an optimal choice, and the evaluation results can be found in Fig. 9(a). In fact, parameter ω_2 is extremely important for our local topology constraint. A large ω_2 easily hinders the saliency transferring procedure, leading to poor detection results. Besides, an appropriate choice of ω_1 is also important for the saliency transferring procedure, and we test several choices of ω_1 with different combination of ω_2. According to the results from Fig. 9(a), we assign $\omega_1 = 10, \omega_2 = 20$ as the optimal choice. Specially, an additional proof toward the superiority of our structure-aware descriptor can be easily observed via comparing the cyan curve (with Structure-aware Descriptor) and the magenta curve (without Structure-aware Descriptor) in Fig. 9 (b).

2) PARAMETER δ

(Eq. 12). We quantitatively evaluate the influences of δ to obtain the optimal choice, and the evaluation results can be found in Table. 1. Actually, the buffer size δ has direct influence on the updating procedure, which is important for the self-adaptive learning procedure to estimate the long-term appearance model. Thus, we assign $\delta = 3$.

TABLE 1. Performance influences from different choices of buffer size δ.

Buffer Size	$\delta = 1$	$\delta = 3$	$\delta = 5$	$\delta = 10$	$\delta = 15$
Precision	0.863	**0.882**	0.862	0.845	0.778
Recall	0.685	**0.746**	0.664	0.676	0.650
F-Measure	0.815	**0.846**	0.807	0.799	0.744

TABLE 2. Performance influences from different choices of SLIC superpixel number.

Superpixel Num	200	300	400	600	800
Precision	0.783	0.868	0.871	**0.882**	0.845
Recall	0.651	0.663	0.710	**0.746**	0.727
F-Measure	0.752	0.812	0.828	**0.846**	0.814

TABLE 3. Performance influences from different choices of the minimum batch size.

Batch Size	7	8	9	10	11	12
Precision	0.876	0.873	**0.882**	0.872	0.867	0.867
Recall	0.747	0.755	**0.746**	0.756	0.758	0.757
F-Measure	0.843	0.842	**0.846**	0.842	0.839	0.839

3) SLIC SUPER-PIXEL SIZE

As shown in Table. 2, it gives rise to remarkable performance improvement by increasing the SLIC super-pixel number at the expense of increased computational cost. However, the quantitative evaluation results indicate that, the optimal super-pixel number is 600. Here it should be noted that, the performance degradation with 800 super-pixels is mainly caused by the absence of the mid-level saliency clues [27], which is the direct consequence of adopting an extremely large super-pixel number.

Specifically, the quantitative result over the minimum batch size (range from 6 to 12) tends to claim ''insensitive'', see details in Table. 3. Although the scores under the different choices are similar to each other, there still exist slight differences when minimum batch size is 9. Thus, we believe these tiny differences are mainly caused by other parameters.

B. QUANTITATIVE EVALUATIONS

In this paper, we evaluate the performance of our method over 4 public benchmarks (almost 200 video sequences), including SegTrack [45], Davis16 [43], DS [44], and UCF [46] dataset. We compare our method with 16 state-of-the-art methods, including FL18 [47], DLVSD18 [48], RADF18 [49], RAS18 [50], DSS17 [51], DHS16 [52], RFCN16 [53], MDF16 [54], DF17 [9], SA15 [39], GF15 [7], MC15 [8], SU14 [12], CS13 [25], HS13 [17], MF13 [18]. To better verify and validate the performance of our method, we leverage the well-recognized precision-recall (PR) as evaluation indicator. We demonstrate the quantitative comparison results in Fig. 10, and the selected qualitative comparisons can be found in Fig. 11.

Due to the lack of global long-term info, the state-of-the-art methods (e.g. DLVSD18 and FD17) easily produce hollow effect, and massive false-alarm detections can be easily observed toward those frame batches with extremely poor low-level predictions. Also, due to the absence of the mechanism to suppress the non-salient surroundings while enhancing the salient foreground, massive false-alarm detections can be easily observed in the detection results of graph-based methods (i.e., GF15, SA15, and MC15). Fortunately, directly benefiting from our novel structure-aware descriptor and low-rank analysis guided fusion strategies, our self-adaptive learning scheme is capable of capturing the long-term common consistencies (i.e, the spatial-temporal coherency) of the foreground object while avoiding the accumulation of false-alarm detections. As for those fusion-based video saliency detection methods (i.e., ST14, and SU14), their massive false-alarm detections are mainly caused by the deficiencies of the sole fusion, which totally neglects the spatial-temporal coherency due to the fact of treating each video frame independently as the naive combination of the motion clues and the color clues. Furthermore, due to the absence of the temporal info, the conventional image saliency methods (e.g., RADF18, RAS18, HS13, and MF13) manifest much worse detections over all adopted benchmarks except for the DS dataset, because the foreground objects in DS dataset frequently show distinct colors against its surroundings. Specifically, since SB14 and BL14 belong to the modeling based methods, which require temporal sequence across a long period of time to construct the robust background model, these methods exhibit good performance for stationary videos but poor performance for non-stationary videos (e.g., the *birdfall* sequence).

C. DIFFERENCES BETWEEN OUR NEW METHOD AND FD17

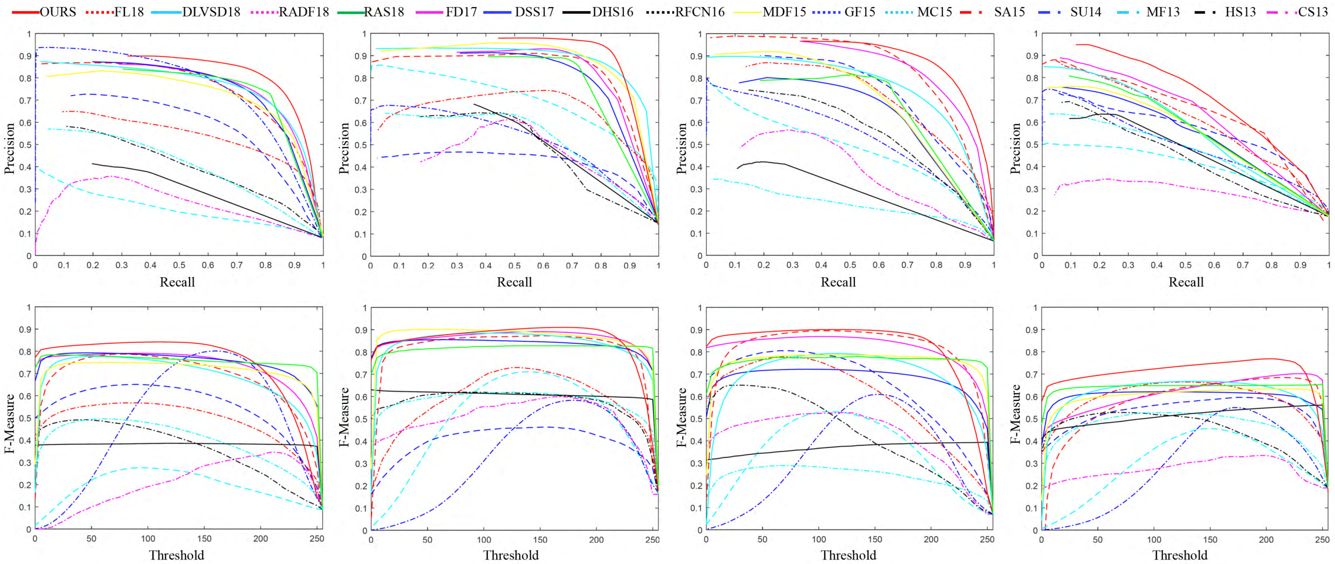
Here we list the theoretical differences between our new method and FD17 [9] in the following two aspects:

First, the feature space adopted by our new method is different to FD17. Since the FD17 only consider the color spanned local info to perform saliency detection, it may easily encounter the obstinate ''error accumulation'' problem if the salient regions and non-salient surroundings exhibit similar color info, see demonstrations in Fig. 6. Benefit from our newly designed non-local descriptor, our new method can well handle the error accumulation problem, see proofs in both quantitative and qualitative comparisons in Fig. 10 and Fig. 11;

Second, the saliency revealing scope of our new method is different to FD17. Although both our new method and FD17 adopted the low-rank analysis for video saliency detection, the low-rank solution adopted in FD17 is designed to guide inter-frame alignments for ''short-term info'' revealing, while we use the low-rank analysis in this paper to facilitate our learning procedure for ''long-term info'' revealing. And the pictorial demonstration of the above differences can be found in Fig. 1. Benefit from the revealed long-term info, our new method can well handle the ''Intermittent Movement Cases'' and ''Fast Scale Change Cases'' while the short-term info based FD17 can not, see comparisons in Fig. 1. And we

TABLE 4. Comparison of quantitative results including MAE (smaller is better) and maximum F-measure (larger is better). The top three results are highlighted in red, green, and blue, respectively.

Metric	maxF	AUC	MAE	maxF	AUC	MAE	maxF	AUC	MAE	maxF	AUC	MAE
DATASET		Davis			DS			SegTrack			UCF	
OUR	0.842	0.968	0.045	0.910	0.957	0.027	0.900	0.987	0.026	0.769	0.927	0.137
FL18	0.739	0.956	0.053	0.833	0.921	0.082	0.831	0.979	0.043	0.636	0.905	0.135
DLVSD18	0.748	0.962	0.055	0.856	0.978	0.057	0.747	0.957	0.046	0.591	0.802	0.152
RADF18	0.781	0.972	0.050	0.888	0.978	0.039	0.807	0.977	0.037	0.606	0.840	0.140
RAS18	0.784	0.919	0.041	0.828	0.907	0.051	0.779	0.789	0.027	0.655	0.692	0.142
FD17	0.758	0.958	0.054	0.845	0.926	0.053	0.820	0.981	0.037	0.645	0.824	0.138
DSS17	0.757	0.963	0.049	0.810	0.934	0.062	0.679	0.876	0.042	0.567	0.787	0.140
DHS16	0.767	0.959	0.043	0.898	0.972	0.042	0.810	0.973	0.035	0.623	0.745	0.143
RFCN16	0.380	0.787	0.082	0.579	0.792	0.094	0.368	0.615	0.063	0.484	0.649	0.180
MDF15	0.720	0.945	0.068	0.859	0.967	0.069	0.714	0.832	0.050	0.576	0.732	0.156
GF15	0.621	0.926	0.099	0.478	0.845	0.111	0.739	0.932	0.078	0.571	0.899	0.146
MC15	0.263	0.728	0.244	0.674	0.873	0.098	0.500	0.928	0.160	0.444	0.825	0.183
SA15	0.554	0.940	0.101	0.716	0.832	0.102	0.716	0.925	0.088	0.601	0.873	0.152
SU14	0.261	0.957	0.102	0.553	0.657	0.122	0.564	0.901	0.125	0.495	0.813	0.167
MF13	0.464	0.878	0.178	0.598	0.733	0.122	0.275	0.807	0.199	0.499	0.802	0.165
HS13	0.455	0.850	0.246	0.591	0.801	0.112	0.601	0.891	0.098	0.479	0.747	0.180
CS13	0.338	0.739	0.108	0.583	0.782	0.118	0.497	0.808	0.127	0.344	0.669	0.186



Davis dataset DS dataset SegTrack dataset UCF dataset

FIGURE 10. Quantitative comparisons (PR curves) between our methods and 15 state-of-the-art methods over, SegTrack [45], [55], Davis16 [43], DS [44] and UCF [46] datasets (almost 200 video sequences).

also believe the above mentioned two cases are quiet common in real video sequences, which shouldn't be ignored.

D. LIMITATIONS

The primary limitation of our method is that, our method tends to be time-consuming in certain sense. Here all the methods are running on a computer with Quad Core i7-4790k 4.0 GHz, 16GB RAM, and NVIDIA GeForce GTX 970. For single 300*300 video frame, although our method has utilized CUDA acceleration, the major bottleneck remains in the pattern model computation (3.61s) and the saliency transferring procedure (10s). It may be noted that, for some cases, high accuracy is perhaps somehow less desirable in the interest of efficiency, so we suggest reducing the SLIC super-pixel number (e.g., reducing from 600 to 400), so that the entire time costs could be reduced by about 60%, and the performance degradation is demonstrated in Fig. 9(e).

VII. CONCLUSION AND FUTURE WORK

In this paper, we have advocated a novel video saliency detection method. Compared with the conventional methods, our method comprises several novel technical components, including: (1) the structure-aware super-pixel based feature descriptor, which automatically enlarge the feature margin while still exhibiting good generalization ability; (2) the self-adaptive learning method, which can captures the common consistence of the salient foreground object in an iterative manner while retaining sufficient adaptability to guarantee

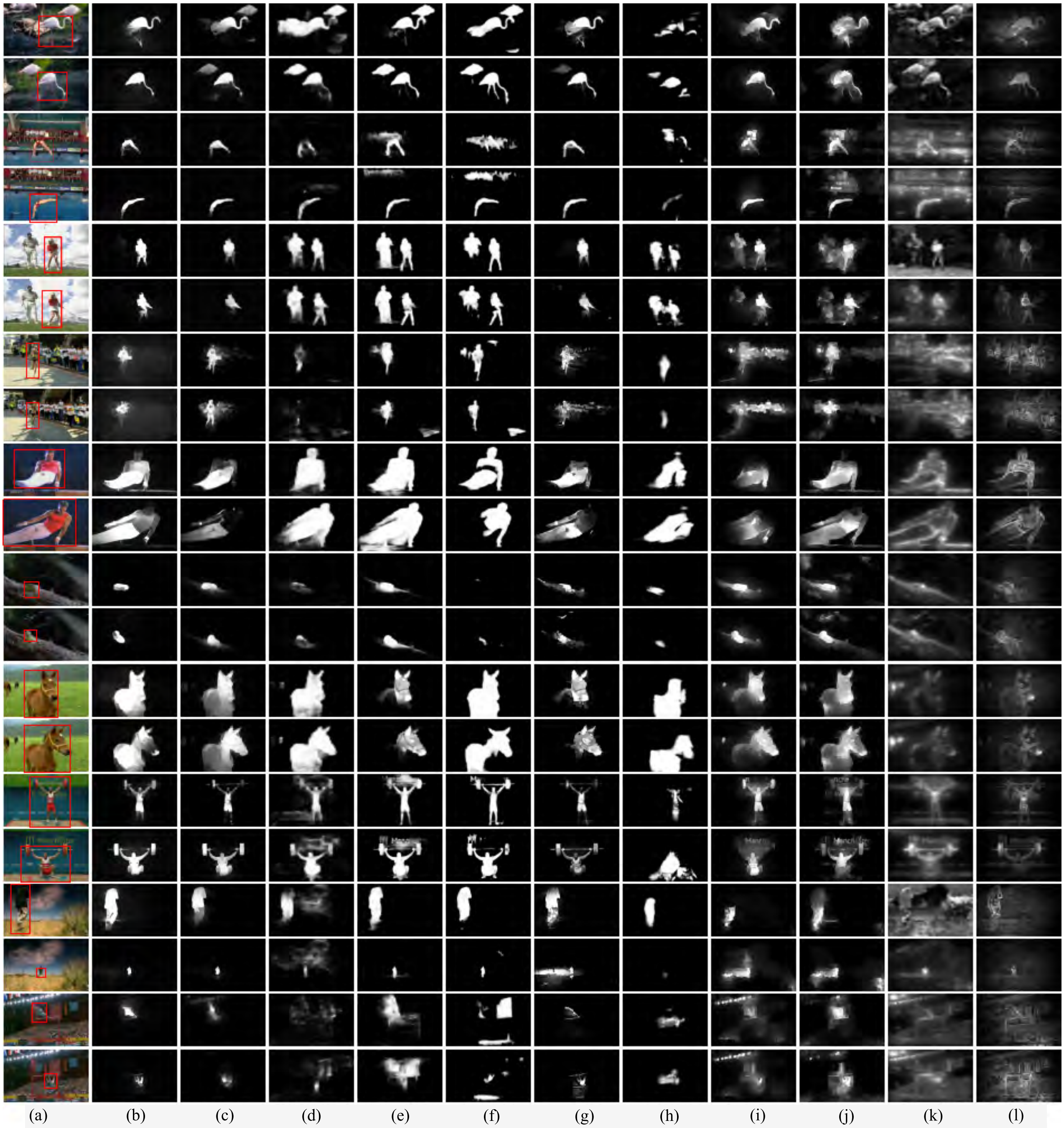


FIGURE 11. Qualitative comparisons over SegTrack [45], Davis16 [43], DS [44] and UCF [46] dataset. (a) denotes the source input image with ground truth marked with red rectangle, (b-l) respectively demonstrate saliency maps of several most representative state-of-the-art methods, including FL18 [47], DLVSD18 [48], RADF18 [49], RAS18 [50], FD17 [9], RFCN16 [53], GF15 [7], SA15 [39], MC15 [8], SU14 [12].

the correct saliency transfer; (3) the low-rank analysis guided selective saliency fusion strategy, which further constrains the global spatial-temporal coherency revealing from the perspective of saliency transfer.

As for our near future work, we are particularly interested in using deep learning solutions to perform fast foregrounds re-identification, which is expected to build an "end-to-end feature tunnel" to align those beyond scope long-term info to the current frames. By applying the low computational spatial-temporal weighting scheme, we may simultaneously achieve good spatial-temporal smoothness while avoiding the obstinate error accumulation limitation. At the same time, to further alleviate the computation cost, our upcoming work will also investigate a much simple solution to perform accurate low-level saliency quality assessment for the anchor frame selection.

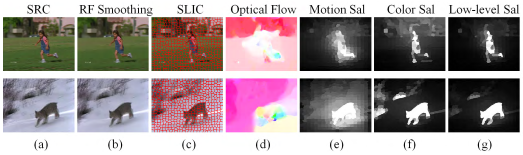


FIGURE 12. Demonstration of the low-level saliency computation.

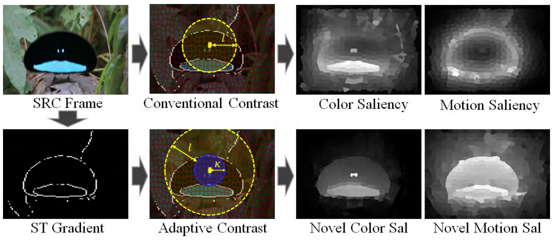


FIGURE 13. Demonstration of the contrast computation.

APPENDIX

LOW-LEVEL SALIENCY COMPUTATION

Given an input video frame $I \in \mathbb{R}^{W\times H\times 3}$, where W and H respectively represents the frame width and height, we adopt the edge-preserving smoothing algorithm **RF** [56] to eliminate unnecessary details of I. Then, the **SLIC** algorithm [57] is used to perform super-pixel decomposition over I to alleviate the contrast computation burden. Meanwhile, we adopt the **Optical Flow** algorithm [58] to capture the motions between consecutive video frames. Here, we denote *vertical flow* as $vx \in \mathbb{R}^{W\times H}$ and *horizontal flow* as $vy \in \mathbb{R}^{W\times H}$. Thus, the *MotionSaliency* can be computed as follows.

$$MS_i = \sum_{p_j\in\psi_i} \frac{||V_i, V_j||_2}{||p_i, p_j||_2}, \quad \psi_i = \{\kappa \le ||p_i, p_j||_2 \le \kappa + l\}, \tag{A1}$$

where $V = [vx \ \ vy]$, $||\cdot||_2$ denotes the l_2-norm, p_i denotes the center position of the i-th super-pixel, ψ_i controls the contrast computation range (the right part of Eq. A1), l is a predefined local contrast computation range used in the conventional local contrast computation, and we empirically set it to $\frac{1}{5}min\{W, H\}$. Also, we formulate the choice of the lower bound κ as:

$$\kappa = \frac{l}{||\Lambda(STG)||_0} \sum_{k\in||k,i||_2\le l} ||\Lambda(STG_k)||_0. \tag{A2}$$

Here, $\Lambda(\cdot)$ denotes the down-sampling operation (30%), *STG* is the spatial-temporal gradient map, whose computation has been demonstrated in Fig. 3 (**ST Gradient**), and the detailed computation can be found in Eq. A3.

$$STG = ||\nabla(I)||_2 \odot ||vx, vy||_2, \tag{A3}$$

where $\odot$ denotes the element-wise Hadamard product, and $\nabla(I)$ denotes color gradient-like contour detections [59]. Obviously, the underlying rationality of our adaptive contrast computation include three aspects: **First**, the conventional contrast computation is easily trapped into hollow effects (**Color/Motion Sal** in Fig. 13). **Second**, the spatial-temporal gradient map is robust enough to capture the structure info of the salient object, because its regional clustering step fully respects the color spatial layout. **Third**, for those super-pixels located at the central region of the salient object, the adoption of κ definitely excludes its surrounding super-pixels from the contrast computation, and finally avoids the hollow effects. Also, the *ColorSaliency* can be computed by Eq. A4.

$$CS_i = \sum_{p_j\in\psi_i} \frac{||(R_i, G_i, B_i), (R_j, G_j, B_j)||_2}{||p_i, p_j||_2}, \tag{A4}$$

where the definition of ψ_i is identical to Eq. A1, and (R_i, G_i, B_i) denote the corresponding averaged *RGB* color of the i-th super-pixel. Thus, the (**Low-level Saliency**) can be easily obtained by multiplying the (**Color Saliency**) with the (**Motion Saliency**) in an element-wise fashion:

$$LS = CS \odot MS. \tag{A5}$$

Here $\odot$ denotes element-wise Hadamard product. As shown in Fig. 12, the fused low-level saliency (**Low-level Sal**) is much better than the motion saliency or the color saliency.

ACKNOWLEDGMENT

(Chenglizhao Chen and Guotao Wang contributed equally to this work.)

REFERENCES

[1] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, “Jointly attentive spatial-temporal pooling networks for video-based person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4743–4752.

[2] S. Varadarajan, P. Miller, and H. Zhou, “Spatial mixture of Gaussians for dynamic background modelling,” in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 63–68.

[3] J.-W. Hsieh, S.-H. Yu, Y.-S. Chen, and W.-F. Hu, “Automatic traffic surveillance system for vehicle tracking and classification,” *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 175–187, Jun. 2006.

[4] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, “Eye guidance in natural vision: Reinterpreting salience,” *J. Vis.*, vol. 11, no. 5, p. 5, 2011.

[5] 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.*, vol. 24, no. 8, pp. 2303–2316, Aug. 2015.

[6] M. Hirokazu, S. Keigo, S. Kazuhito, and S. Nobuhiro, “Basic design of visual saliency based autopilot system used for omnidirectional mobile electric wheelchair,” *Comput. Sci. Inf. Technol.*, vol. 3, no. 5, pp. 171–186, 2015.

[7] W. Wang, J. Shen, and L. Shao, “Consistent video saliency using local gradient flow optimization and global refinement,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[8] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, “Spatiotemporal saliency detection for video sequences based on random walk with restart,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015.

[9] 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.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[10] S. H. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, “Video saliency detection via dynamic consistent spatio-temporal attention modelling,” in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1063–1069.

[11] 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 Process. Lett.*, vol. 25, no. 2, pp. 154–158, Feb. 2018.

[12] Y. Fang, Z. Wang, W. Lin, and Z. Fang, “Video saliency incorporating spatiotemporal cues and uncertainty weighting,” *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.

[13] F. Zhou, S. B. Kang, and F. C. Michael, "Time-mapping using space-time saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3358–3365.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[15] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[16] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-Y. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2976–2983.

[17] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[18] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[19] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[20] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[21] J. Li, L.-Y. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the secret of image saliency in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2428–2440, Dec. 2015.

[22] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[23] J. Yan, M. Zhu, H. Liu, and Y. Liu, "Visual saliency detection via sparsity pursuit," *IEEE Signal Process. Lett.*, vol. 17, no. 8, pp. 739–742, Aug. 2010.

[24] 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.*, Jun. 2012, pp. 733–740.

[25] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[26] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–12.

[27] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.

[28] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[29] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, 2009.

[30] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.

[31] Y. Li, Y. Tan, J. Yu, S. Qi, and J. Tian, "Kernel regression in mixed feature spaces for spatio-temporal saliency detection," *Comput. Vis. Image Understand.*, vol. 135, no. 1, pp. 126–140, 2015.

[32] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[33] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[34] Z. Gao, L.-F. Cheong, and Y.-X. Wang, "Block-sparse RPCA for salient motion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1975–1987, Oct. 2014.

[35] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[36] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[37] 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.*, vol. 52, pp. 410–432, Apr. 2016.

[38] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 628–635.

[39] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3395–3402.

[40] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[41] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approx.*, vol. 28, no. 3, pp. 253–263, Dec. 2008.

[42] D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *J. Comput. Syst. Sci.*, vol. 66, no. 4, pp. 671–687, 2003.

[43] F. Perazzi, J. Pont-Tuset, 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.*, Jun. 2016, pp. 724–732.

[44] F. Ken, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun./Jul. 2009, pp. 638–641.

[45] F. Li, T. Kim, A. Humayun, D. Tsai, and M. R. James, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2192–2199.

[46] M. Stefan and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 842–856.

[47] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, "Bilevel feature learning for video saliency detection," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3324–3336, Dec. 2018.

[48] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[49] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. Assoc. Adv. Artif. Intell.*, 2018, pp. 6943–6950.

[50] S. Chen, X. Tan, B. Wang, and X. Hu. (2018). "Reverse attention for salient object detection." [Online]. Available: https://arxiv.org/abs/1807.09940

[51] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5300–5309.

[52] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 678–686.

[53] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 825–841.

[54] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5455–5463.

[55] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.

[56] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–12, Jul. 2011.

[57] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels," EPFL, Lausanne, Switzerland, Tech. Rep. 149300, 2010.

[58] C. Liu, "Exploring new representations and applications for motion analysis," Ph.D thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[59] M. Leordeanu, R. Sukthankar, and C. Sminchisescu, "Efficient closed-form solution to generalized boundary detection," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 516–529.