Bi-level Feature Learning for Video Saliency Detection

Chenglizhao Chen[1] Shuai Li[2] Hong Qin[3] Zhenkuan Pan[1*] Guowei Yang[4]
[1]College of Computer Science and Technology, Qingdao University
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[3]Stony Brook University
[4]College of Electronic Information, Qingdao University

***Abstract*—This paper advocates a novel learning solution to the modeling of long-term spatial-temporal saliency consistency in order to boost the accuracy for video saliency detection. Conventional methods typically utilize the "slack" spatial-temporal model to locally ensure the smoothness of the computed video saliency, yet they could easily encounter the performance trade-off dilemma (i.e., detection' accuracy and integrity). In contrast, our novel approach proposes the bi-level learning strategy to globally exploit the saliency consistency while overcoming the above difficulty. Our method first starts with the contrast computation of low-level saliency clues in a frame-wise manner. Then, based on such obtained saliency clues, we devise a novel bi-level Markov Random Field (bMRF) solution to conduct semantic labelling, which can explicitly indicates both the salient salient foregrounds and non-salient nearby surroundings with high confidence while shrinking the low confidence remains. In such a way, the spatial-temporal consistency constraint is embedded intrinsically into the above explicit semantic labels, and we avoid the performance trade-off problem from occurring. Next, based on those semantic labels made by our bMRF method, we further propose to learn multiple non-linear feature transformations to enlarge the feature margin between the salient foregrounds and the non-salient nearby surroundings, whose key rationale is to resort to long-term common consistencies to enforce the spatial-temporal smoothness. Thus, we can utilize these learned non-linear feature transformations to simultaneously suppress those short-term false-alarms and correct those hollow effects. To validate our new approach, we conduct extensive experiments on 5 publicly available benchmarks, and make comprehensive, quantitative evaluations between our method and 17 state-of-the-art techniques. All the results demonstrate our method's advantages in terms of accuracy, reliability, robustness, and versatility.**

***Index Terms*—Spatial-temporal Saliency Consistency, Bi-level Morkov Random Field, Localized Feature Transformation, Video Saliency Detection.**

I. INTRODUCTION AND MOTIVATION

THE main purpose of video saliency detection is to indicate the given video sequence' most salient object, which frequently exhibits uniqueness movement pattern. As a pre-processing tool, the video saliency detection currently has become a critical factor to dictate the performance of the corresponding downstream applications, including visual tracking [1], [2], space-time visualization [3], video resizing [4], and video compression [5] [6]. After years of extensive studies, researchers have reached a consensus that the key factor toward the robust video saliency detection relies on two aspects: the low-level saliency computation [7] [8] and the usage of spatial-temporal consistency to boost the detection accuracy [9] [10]. In fact, since the first problem has been well studied in our previous works [7] [11], the main foci of this paper is to further boost the accuracy and robustness of the detected video saliency by proposing a novel strategy to explore the saliency consistency simultaneously from spatial and temporal perspectives. As for the most-represented solution to explore the saliency consistency from the spatial-temporal perspective, the graph based methods simultaneously integrate both the spatial and the temporal info into an unified energy framework to iteratively perform two steps: the iter-frame alignment step, then the spatial-temporal saliency weighting. Although the spatial-temporal saliency consistency degree can be implicitly revealed in "short-term" manner by performing the graph based energy minimization, several obstinate difficulties still remain unsolved, and we shall give a brief analysis about these challenges.

First, the usage of the spatial-temporal consistency toward the saliency boosting could easily encounter the performance trade-off, i.e., an overemphasis on the spatial-temporal consistency definitely damages the detection' integrity , and vice versa. So, the bottleneck is obvious that an extremely strong consistency constraint is indispensable for those poor quality low-level saliency clues to retain spatial-temporal smoothness, however, such aggressive implementation inevitable to enhances those false-alarm detections and finally causes the false-alarm accumulations.

Second, the saliency exploration scope of the conventional graph methods are frequently too local to obtain robust video saliency, especially for those short-term dynamic background [12] caused false-alarm detections. For example, the graph energy solutions [13] [14] only utilize the *beyond scope* spatial-temporal saliency information to re-boost the accuracy of current frame's saliency prediction. However, since the conventional graph methods treat all the temporally neighbored frames equally via the *majority voting* like scheme, it is difficult to resort the short-term spatial-temporal saliency consistency alone to obtain integral saliency detection while delimitating those false-alarm detections.

To tackle the aforementioned first challenge (i.e., the performance trade-off), our current research efforts are endeavored to take full advantage of the graph model to exploit the

Corresponding author: Zhenkuan Pan (zkpan@qdu.edu.cn)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

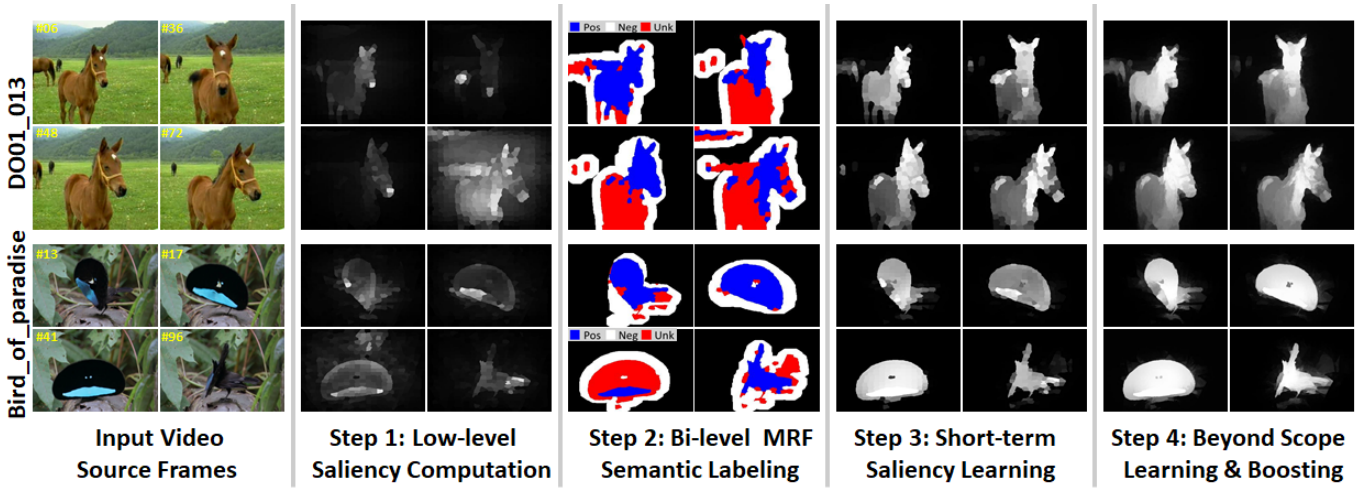

Fig. 1. The architectural overview of our video saliency detection method. Our method first utilizes the contrast solution to compute the low-level saliency, which are already demonstrated in Step 1. And then in Step 2, our method automatically regards the intersections of the two-level MRF saliency assumption as the most trustworthy **Pos** region (which means the salient regions are denoted by blue color), and regards the near surrounded outliers of these two-level assumptions union as the most trustworthy **Neg** region (which means the non-salient regions are denoted by white color). Thus, based on the bi-level saliency assumptions, i.e., the positive **Pos** and negative **Neg** instances, our learning scheme can automatically assign saliency value to those undeterministic regions **Unk** in Step 3. The saliency detection results of Step 4 are obtained by performing our long-term learning scheme with superpixel-wise and pixel-wise spatial-temporal smoothing schemes.

short-term spatial-temporal consistency while avoiding the performance trade-off dilemma by incorporating the bi-level Markov Random Field (bMRF) strategy.

That is, we equally divide the input video sequence into short-term video frame batches (identical to our previous work [7]) and then build the spatial-temporal graph (Fig. 3c) batch-wisely to integrate both the spatial and temporal info. In fact, based on the above spatial-temporal graph, the conventional MRF is the most intuitive choice to reveal short-term spatial-temporal saliency consistency, and the consistency degree is determined by the MRF' smoothness term. That is, the more we emphasize on the MRF' smoothness term in energy minimization, the obtained binary assumption will exhibit stronger spatial-temporal saliency consistency. Thus, in order to strike the conventional MRF' performance trade-off (i.e., trade-off between accuracy and integrity), our bMRF respectively output two-level of binary assumptions, i.e., one level with strong saliency consistency which sacrifices the assumption' integrity (we name it "aggressive" saliency assumption), another level with relatively weak saliency consistency while exhibiting better assumption' integrity (we name it "conservative" saliency assumption). Therefore, the intersections (**Pos** in Fig. 1) between these two-level saliency assumptions frequently corresponds to the most confident salient regions, while the near surrounded outliers of these two-level assumptions' union (**Neg** in Fig. 1) always corresponding to the most trustworthy non-salient backgrounds. In this way, the spatial-temporal saliency consistency constraint is now intrinsically embedded into the above explicit semantic labels. Specially, we propose to use our newly designed metric learning strategy to automatically determine the remaining regions (**Unk** in Fig. 1) via enlarging the distance margin between the trustworthy salient and non-salient assumptions.

As for the second challenge (i.e., long-term saliency modeling), we propose to extend the learning scope of our novel learning strategy in a batch-wise fashion, whose behind rational is to resort the long-term common consistency to either eliminate current false-alarms or fill the long period hollow detections (Fig. 2). That is, for each video frame batch, we learn multiple non-linear feature transformations from current low-level saliency to capture the intra-batch's video saliency, and all these learned transformations are pooled together to ensure the long-term smoothness of the computed video saliency. Meanwhile, in order to avoid the over-fitting problem, we propose to reduce the learning ambiguity via constraining our metric learning extent within non-overlapped local constraint pairs. To summarize, the main contributions of this paper are two-fold:

- We propose a bi-level Markov Random Field (bMRF) method to strike the performance trade-off between the integrity and the accuracy within batch-wise manner. Thus, by using our bMRF method, our subsequent metric learning can take the full advantage of the spatial-temporal consistency to boost the accuracy of video saliency detection.
- We propose a localized metric learning solution to implicitly reveal the long-term spatial-temporal saliency consistency. Thus, our novel learning solution can automatically fill those long period hollow effects (Fig. 2) while eliminating intermediate false-alarm detections.

II. BACKGROUND AND RELATED WORKS

In fact, the rationale of the spatial-temporal coherency guided video saliency methods is that, the movement trajectories of the salient foreground object are frequently characterized by spatial-temporal smoothness. Thus, long-term modeling/learning with appropriate update is the most intuitive solution to exploit the spatial-temporal coherency in consecutive video frames, e.g., the deep learning methods [15] [16]. From the perspective of scene modeling, background subtraction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

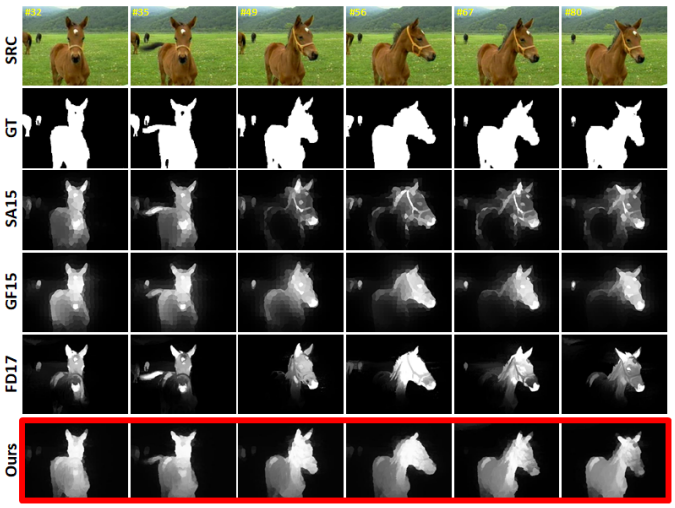

Fig. 2. Qualitative comparisons between the state-of-the-art methods (GF15 [10], SA15 [9], FD17 [7]) and our method towards the long period stop-and-go movement. Apparently, our method can well utilize the learned long-term feature transformations to fill the hollow effects of the standstill horse.

based salient motion/change detection methods [17], [18], [19] have been well studied in recent years, whose central idea is to utilize low-rank decomposition [20] 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 [19] due to the absence of pixel-wise correspondence in consecutive video frames. To ameliorate, either frame-level affine registration [21] or background tracking strategy [22] 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 impact 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, [23] proposes to utilize the foreground spatial-temporal information to construct their attention model, which fully takes the advantage of the motion continuity to eliminate false-alarm detections. Similarly, Li et al. [24] proposed to utilize the newly-designed kernel regression to exploit 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. [13] regarded the graph model 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 to retain spatial-temporal consistency. So, [25] 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. [10], [9]. Although these graph model based video saliency methods have achieved remarkable performance improvement, the graph model solution easily causes the accumulation of false-alarm errors, because its un-bounded saliency expansion lacks of a mechanism to suppress the non-salient backgrounds while enhancing the salient foreground object.

Although the most recent method [7] utilizes the low-rank strategy to alleviate the problem of false-alarm error accumulation in a batch-wise manner, it does not take full advantage of the beyond scope saliency consistency, which easily produces massive false-alarm detections when the majority of the intra batch's saliency clues are incorrect. Therefore, this paper will exploit the long-term spatial-temporal coherency in a learning fashion to avoid the above problems while ensuring the spatial-temporal smoothness of the computed video saliency.

III. METHOD OVERVIEW

Algorithm 1. Main Steps of Our Video Saliency Method

Initialization:

Perform RF [26] smooth to each video frames;
Perform Optical Flow [27] to sense motion;
Perform batch decomposition [7] (frame number$\geq$9);
Perform SLIC [28] over-segmentation;
Perform multi-level feature representation (Sec. V-A);
Initialize the feature transformations $\mathbf{A} = \mathbf{I}$;

For each video frame batch

1. Perform contrast based low-level saliency (Sec. IV-A);
2. Construct spatial-temporal graph based on **LS** (Fig. 3c);
3. Obtain bi-level saliency assumptions (Sec. IV-B);
4. Learn non-linear feature transformations **A** (Sec. V-B);
5. Compute learned video saliency (Sec. V-D);

End For

As we can see in Fig. 1, our video saliency detection method mainly consists of four components: (1) low-level saliency computation; (2) bi-level semantic labelling; (3) short-term video saliency learning; and (4) beyond-scope saliency learning and boosting (i.e., spatial-temporal saliency smoothing and pixel-wise saliency assignment). Our method first computes the low-level video saliency frame-by-frame and then equally decomposes the long-term source sequence into short-term frame batches. After that, based on the pre-obtained low-level saliency clues, we propose the bi-level Markov Random Field (bMRF) model to perform batch-wise saliency assumption, which can automatically identify those trustworthy salient/non-salient regions via semantic labeling. Then, according to the above bMRF saliency assumptions, we adopt the localized learning solution to enlarge the feature margin between the salient regions (blue **Pos**) and the non-salient surroundings (white **Neg**), and those undeterministic regions (red **Unk**) are excluded from our learning iterations to maintain Generalization ability. Therefore, benefiting from the non-salient surrounding's feature coherency and our newly proposed beyond-scope learning scheme, those undetermined nearby surroundings (i.e., red **Unk** in Step 2 of Fig. 1) can be assigned with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

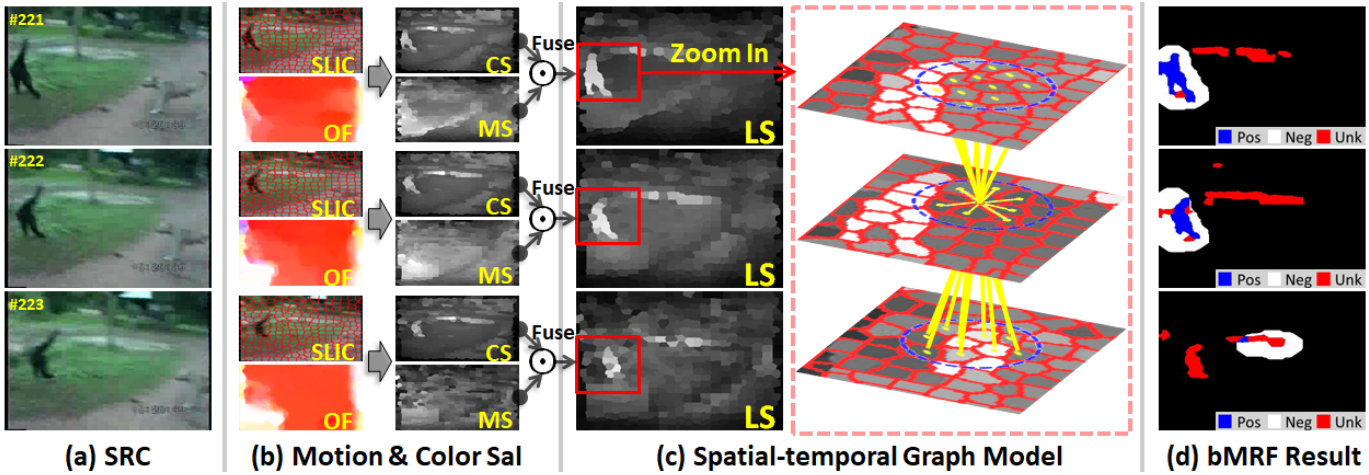

Fig. 3. Demonstration of our graph model construction, where the right-bottom yellow marks respectively denote the SLIC superpixel decomposition (SLIC), the Optical Flow computation (OF), the Motion Saliency (MS), the Color Saliency (CS) and the Low-level Saliency (LS). The yellow dots in the graph model (c) denote the center position of superpixels (i.e., graph' nodes), and the yellow lines represent the color distance (i.e., graph' edges) to measure the similarity between two neighbored graph nodes. Since the foreground mask **FM** is determined within the batch-wise manner, the obtained foreground masks (i.e., white regions in d) may exist large difference between different frame batches. Also should be noted that our bMRF method may exists incorrect saliency assumption, e.g., the blue superpixel in bottom row of d.

appropriate saliency values. Meanwhile, we adopt a series of saliency boosting strategies (e.g., spatial-temporal smoothing and pixel-wise saliency assignment) to further maintain the spatial-temporal consistency of the learned video saliency.

For better understanding, we summarize all the necessary steps of our video saliency method in Algorithm 1.

IV. BI-LEVEL SALIENCY ASSUMPTION

A. Low-level Video Saliency

Different from the conventional image saliency detection which relies the spatial info only, the incursion of temporal info is the critical factor for the correct video saliency detection. For any input video sequence, the most salient region is simultaneously determined by its corresponding difference toward its surrounding in both spatial extent and temporal scale. Thus, we propose to utilize the contrast computation over the SLIC [28] over-segmented mid-level feature space (column b in Fig. 3) to explore those saliency clues, i.e., the color saliency and motion saliency. That is, we propose to compute the motion saliency **MS** in Optical Flow [27] gradient spanned feature space (i.e., two-direction gradients: vx and vy), and compute the color saliency **CS** in RGB color space, which can be detailed as follows:

$$\mathbf{MS}_i = \sum_{\mathbf{p}_j \in \psi_i} \frac{||\mathbf{V}_i, \mathbf{V}_j||_2}{||\mathbf{p}_i, \mathbf{p}_j||_2}, \quad \mathbf{CS}_i = \sum_{\mathbf{p}_j \in \psi_i} \frac{||C_i, C_j||_2}{||\mathbf{p}_i, \mathbf{p}_j||_2}, \tag{1}$$

where C_i represents averaged RGB value of the i-th superpixel, $\mathbf{V} = [vx \quad vy]$, $||\cdot||_2$ denotes the l_2-norm, $\mathbf{p}_i$ denotes the center position of the i-th super-pixel, ψ_i controls the contrast computation range: $\psi_i = \{||\mathbf{p}_i, \mathbf{p}_j||_2 \leq 200\}$. Also, we conduce the *min-max* normalization batch-wisely to produce final motion saliency and color saliency.

$$\mathbf{LS} = \xi(\mathbf{CS}) \odot \xi(\mathbf{MS}). \tag{2}$$

Since the above saliency clues are independently computed in frame-by-frame manner, those false-alarm detections can be further filtered by seeking saliency coherency between consecutive video frames. Therefore, we propose to use the spatial-temporal weighting scheme to further boost the accuracy of the above obtained saliency clues, and then fused it as the low-level saliency **LS** in an element-wise fashion (Eq. 2).

$$\xi(\mathbf{MS}_i^t) : \mathbf{MS}_i^t \leftarrow \frac{\sum_{p=t-1}^{t+1} \sum_{q \in \psi} e^{-\theta \cdot ||C_i^t - C_q^p||_2} \times \mathbf{MS}_i^t}{\sum_{p=t-1}^{t+1} \sum_{q \in \psi} e^{-\theta \cdot ||C_i^t - C_q^p||_2}}. \tag{3}$$

Here $\xi(\cdot)$ denotes the spatial-temporal weighting function, C_i^t represents RGB value of the i-th superpixel in t-th video frame, $q \in \psi$ constraints the spatial weighting' computation range which we empirically assign 25 as the maximum Euclidean radius, θ controls the color discriminative power that we empirically assign it to 30 to follow the suggestion of [11], $\odot$ denotes element-wise Hadamard product. Obviously from Fig. 3c that the low-level saliency **LS** is much better than either the motion saliency **MS** or the color saliency **CS**, and the quantitative proofs can be found in Fig. 14.

B. Bi-level Markov Random Field Guided Saliency Assumption

Based on the low-level saliency clues, which are computed via frame-wise manner in previous subsection, we propose to utilize the spatial-temporal consistency to further boost the accuracy of the above low-level saliency, i.e., enhance the common consistency of the salient foregrounds while compressing those non-salient remains. Since the graph based solutions can well reveal the short-term spatial-temporal consistency, we propose to utilize Markov Random Field (MRF), which is one of the most representative graph solution, to perform binary saliency assumption in batch-wise fashion, and the formulation of the spatial-temporal graph can be found in Fig. 3c. In fact, the conventional Markov Random Field (MRF) model consists two components, i.e., the data term

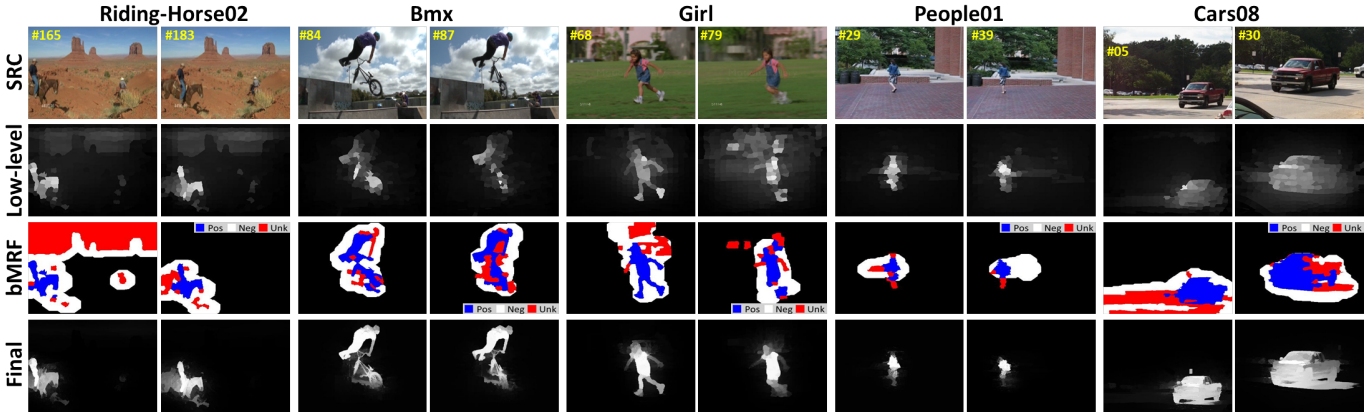

Fig. 4. Demonstration of the learned video saliency toward our bi-level Markov Random Field (bMRF) saliency assumptions. The first row represents the input video source frames, the second row demonstrates the low-level saliency, the third row is our bMRF saliency assumption results, and the last row demonstrates the learned video saliency. Apparently, those undetermined regions (red regions in third row) can be correctly determined by our learning solution.

(i.e., the left part of Eq. 4) and the smoothness term (i.e., the right part of Eq. 4), thus the energy minimization procedure of MRF (GCO toolbox [29]) is sensitively controlled by the strength parameters (i.e., ε in Eq. 4 and β in Eq. 5). Following the "one-fix-all" rational, the main foci of the previous MRF related works generally relies on the formulation of the exact optimal MRF strength parameter, which attempt to suit various scenarios with one fixed strength parameter, i.e., we name it as single level MRF. However, due to the variation nature of salient foregrounds, it is difficult for the single level MRF to satisfy the long-term video saliency detections. Therefore, to combat this limitation, we propose to adopt the bi-level Markov Random Field (bMRF) model to convert the conventional binary saliency assumption problem into multi-layer semantic labelling. That is, instead of the conventional "one-fix-all" strategy, we propose to respectively use the "aggressive" strength parameters (ε_a in Eq. 4 and β_a in Eq. 5) to focus on the detection' accuracy and use the "conservative" strength parameters (ε_c in Eq. 4 and β_c in Eq. 5) to bias toward the detection' integrity. Here we formulate our bMRF solution as Eq. 4, and we use the subscript a/c to respectively denote the strength parameters of the "aggressive" level MRF and "conservative" level MRF. Also, we adopt an additional constraint to enhance the consistency of inter-batch' bMRF binary assumptions, see details in Eq. 4.

$$\begin{aligned} &\min_{S_{a/c}} \sum_i u(S_i) + \lambda \sum_{i,j \in \varepsilon_{a/c}} e^{-\theta \cdot |C_i - C_j|} \cdot |S_i - S_j|, \\ &s.t. \quad |\Lambda(||S^p||_0) - \Lambda(||S^q||_0)| \leq \alpha \cdot \xi, \end{aligned} \tag{4}$$

where $S \in \{0, 1\}$ represents the binary saliency assumption, $u(\cdot)$ is an unary function (Eq. 5), parameter β controls the bias tendency toward the data term, **LS** denotes the low-level saliency, parameter λ controls the strength of the smoothness term and we assign it to 1 to follow the suggestion of [21], θ controls the color discriminative power, which is identical to θ in Eq. 3, ε represents the spatial-temporal neighborhood regions which is demonstrated as the blue dash circle in Fig. 3c, the average function $\Lambda(\cdot)$ only considers those l_0-norms between 20-th and 80-th percentiles, $\alpha = 0.1 \times \Lambda(||S^p||_0)$, p and q are respectively from consecutive frame batches, and ξ denotes the slack variable.

$$u(S_i) = \begin{cases} 1, & if \ (\mathbf{LS}_i \cdot S_i) \geq \beta_{a/c} \cdot std(\mathbf{LS})^2 \\ 0, & otherwise \end{cases}. \tag{5}$$

In fact, Eq. 4 can be effectively solved via iteratively performing the convex optimization and adjusting $\beta_{a/c}$ to satisfy the temporal constraint. Specially, we formulate the updating procedure of $\beta_{a/c}$ as Eq. 6. Meanwhile, to guarantee the convergency, we also shrink the slack variable $\xi \leftarrow 0.9 \times \xi$ when the residual' $sign$ change, and we assign the initial $\xi = 1$.

$$\beta_{a/c} \leftarrow \beta_{a/c} \cdot (1 + 0.1 \times sign(\Lambda(||S^p||_0) - \Lambda(||S^q||_0)) \cdot \xi). \tag{6}$$

So far, we can obtain the bi-level saliency assumptions via respectively assigning $\{\beta_a, \varepsilon_a\}$ and $\{\beta_c, \varepsilon_c\}$ to Eq. 4. As $\beta_{a/c}$ inversely controls the performance trade-off toward the $\varepsilon_{a/c}$, we empirically assign $\varepsilon_a = 30, \varepsilon_c = 45$, and then perform quantitative evaluation to obtain the optimal choice of $\beta_{a/c}$ (details can be found in Sec. VI-A). And we also define the foreground mask **FM**, which comprises all possible salient regions, as $d(N(\sum S_c) > 0.15)$, where $d(\cdot)$ denotes the dilation operator with 20×20 gaussian mask, N represents the *min-max* normalization. Specially, the rational behind of the foreground regions is to utilize the inter-frame' coherency to coarsely locate regions containing all possible salient foregrounds.

Based on the above computed two-level binary saliency assumptions (i.e., S_a and S_c), we can perform the semantic labelling (**Step 2** in Fig. 1) as following:

(1) we regard the intersections of the obtained two-level saliency assumptions (i.e., $(S_a \cap S_c)_+$) as those most trustworthy salient foreground regions, see **Pos** in Fig. 5.

(2) we regard the residual between the foreground mask **FM** and the union of two-level binary saliency assumptions (i.e., $\{\mathbf{FM} - (S_a \cup S_c)_+\}_+$) as those most trustworthy non-salient background regions, see **Neg** in Fig. 5.

(3) we regard the positive residual between the union of two-level binary saliency assumptions and **Pos** (i.e., $\{(S_a \cup S_c)_+ - (S_a \cap S_c)_+\}_+$) as those undetermined regions (**Unk** in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

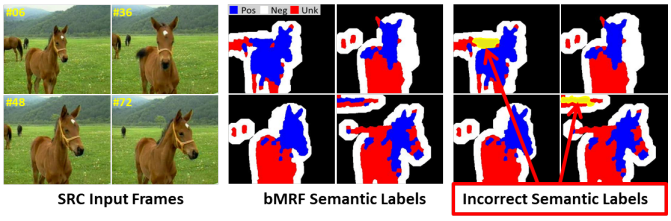

Fig. 5. Demonstration of those incorrect semantic labels, which is mainly caused by exceptions of the bMRF' strength parameters.

Fig. 5), which we propose to determine it via our subsequent metric learning solution.

That is, by regarding bMRF determined semantic labels as training instances (Fig. 3d), we can resort metric learning solution, which will be further introduced in Sec. V, to estimate an appropriate decision boundary for the saliency assignment toward those undetermined remaining regions (**Unk**).

Here we summary the main differences between the conventional MRF and our newly proposed bMRF method in three aspects:

First, the graph structure of our bMRF is slightly different to the conventional MRF (see the $\varepsilon_{a/c}$ in Eq. 4). That is, we use varying strength parameters (i.e., $\varepsilon_{a/c}$ in Eq. 4 and $\beta_{a/c}$ in Eq. 5) to facilitate our two-level saliency assumptions (i.e., the "aggressive" level and the "conservative" level).

Second, since the short-term saliency consistency is estimated in batch-wise manner, we put an additional long-term constraint (see the constraints in Eq. 4) to ensure the iter-batch smoothness to further robust the short-term saliency consistency revealing.

Third, the subsequent usage of our bMRF is also different to the conventional MRF method. As for the conventional MRF methods, the estimated video saliency is solely related to the MRF' binary saliency assumption. However, our bMRF solution simultaneously considers both bi-level assumptions to formulate implicit semantic labels, whose quality directly determines the performance of our subsequent learning procedure.

V. LONG-TERM VIDEO SALIENCY MODELING

A. Feature Representation

Due to the low discriminative power of the RGB color space, it is difficult to resort learning solution directly to obtain reasonable discriminative margin, thus we propose to integrate multi-scale spatial info. That is, for each SLIC over-segmented superpixel (total superpixel number is 500 in single video frame), we formulate our feature space F as follows:

$$F \in \mathbb{R}^{1\times 45} = \{RGB^3, Lab^3, CN^3, \\ RGB^4, Lab^4, CN^4, RGB^5, Lab^5, CN^5\}, \tag{7}$$

where the up-script $3, 4, 5$ respectively denote the features obtained from different scale, i.e., SLIC over-segmentation with total superpixel number 300, 400 and 500 respectively, $CN \in \mathbb{R}^{1\times 10}$ [30] represents the color mapping results, which is computed by converting the original 3 dimensional RGB color into 10 dimension linguistic color labels. Specially, we only consider the last two channels of Lab color info.

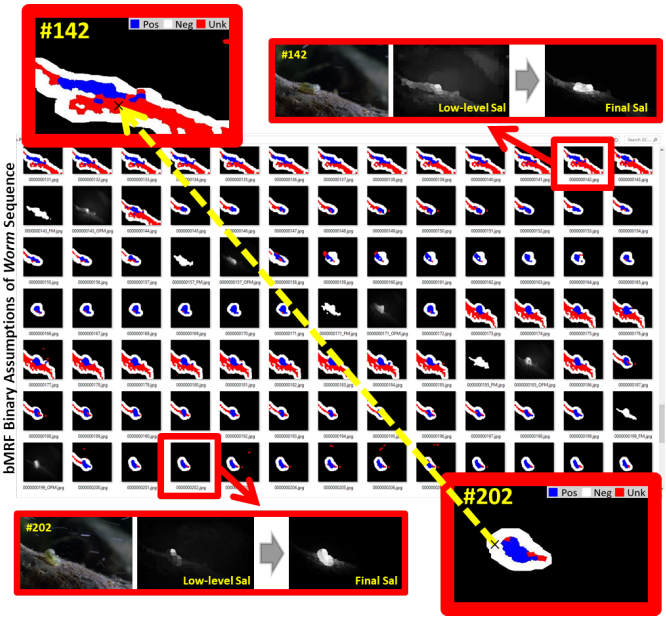

Fig. 6. Demonstrate of the intuition of our novel method toward the rejection of non-salient regions. Benefit from these learned beyond scope info, i.e., the long-term saliency consistency (**Neg** in #202), our method can automatically determine those previous undetermined regions (**Unk** in #142), see the yellow dash arrow.

Thus, the dimension of feature space F can be detailed as by $(RGB \in \mathbb{R}^{1\times 3} + Lab \in \mathbb{R}^{1\times 2} + CN \in \mathbb{R}^{1\times 10}) \times 3 = 45$.

B. Localized Saliency Learning

Based on the batch-wise semantic labels made by our bi-level Markov Random Field (bMRF) method, we propose to learn multiple non-linear feature transformations $\mathbf{A} \in \mathbb{R}^{45\times 45}$ to enlarge the feature distance ($d_A(F_{Pos}, F_{Neg})$ in Eq. 8) between the salient foregrounds (**Pos** with total m instances) and the non-salient near surroundings (**Neg** with total n instances).

$$d_A(F_{Pos}, F_{Neg}) = (F_{Pos} - F_{Neg})^T \mathbf{A}(F_{Pos} - F_{Neg}). \tag{8}$$

Thus, by regarding the **Neg** as the contrast anchors, we can estimate the learned saliency degree of those undetermined regions (**Unk**) via contrast computation. That is, we automatically assign appropriate saliency value to those undetermined regions (**Unk**) according to the transformed feature distance (Eq. 9).

$$d_A(F_{Neg}, F_{Unk}) = (F_{Neg} - F_{Unk})^T \mathbf{A}(F_{Neg} - F_{Unk}). \tag{9}$$

So far, the above learning problem can be relaxed into satisfying implicit feature distance constraints while conserving unbiased $\mathbf{A}$ to avoid over-fitting problem (i.e., ensuring the closeness between $\mathbf{A}$ and identity matrix $\mathbf{I}$), and the detailed learning formulation can be found in Eq. 10.

$$\begin{aligned} &\min_{\mathbf{A}\succeq 0} \ ||\mathbf{A}, \mathbf{I}||_2 \\ &s.t. \ tr(\mathbf{A}(F_i - F_j)(F_i - F_j)^T) \geq \xi_l \ i \in \textbf{Pos} \ j \in \textbf{Neg}, \\ &\quad tr(\mathbf{A}(F_i - F_j)(F_i - F_j)^T) \leq \xi_u \ (i,j) \in \textbf{Pos}, \\ &\quad tr(\mathbf{A}(F_i - F_j)(F_i - F_j)^T) \leq \xi_u \ (i,j) \in \textbf{Neg}, \end{aligned} \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

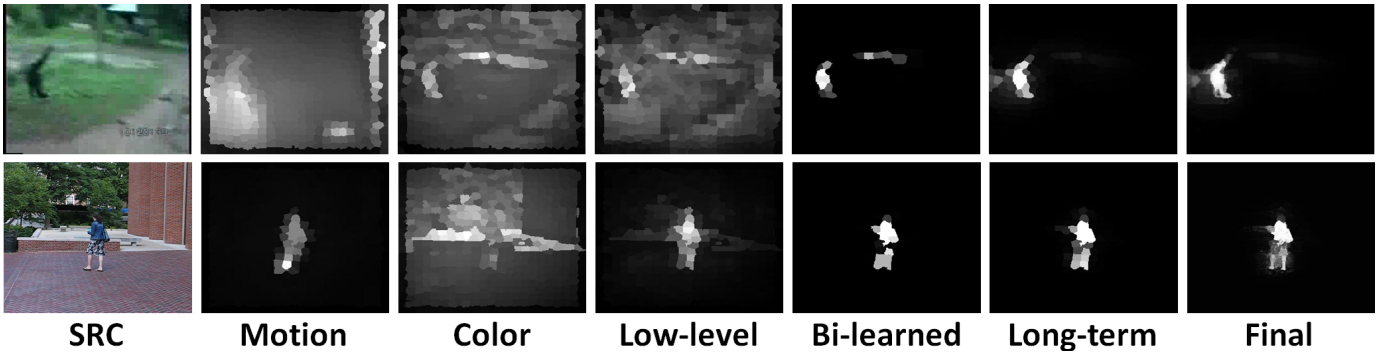

Fig. 7. The qualitative demonstration via combining different components of our method, where the **Motion** column demonstrates the motion saliency map, the **Color** column demonstrates the color saliency map, **Low-level** illustrates the motion and color fused low-level saliency map (Eq. 2), **Bi-learned** illustrates the obtained saliency map after introducing our bMRF guided metric learning, **Long-term** illustrates the obtained saliency map after introducing the beyond scope learned feature transformation, and **Final** denotes the final video saliency detection after spatial-temporal pixel-wise saliency assignment. Also, the corresponding quantitative evaluation result can be found in Fig. 14.

where ξ_u and ξ_l represent the slack variables, and we initialize ξ_u/ξ_l as the upper/lower bound of the constraints' l_2 feature distance. Apparently, the rational behind the explicit constraints in Eq. 10 is to enlarge the "*inter class*" feature distance while shrinking the "*intra class*". And we can obtain $m \times n$ constraints, which is consisted by instances respectively from **Pos** and **Neg**, to enlarge the feature margin between two different class.

However, due to the occasional existence of the incorrect bMRF saliency assumption (see demonstrations in Fig. 5), it could easily bias the learning problem into local minima if we attempt to satisfy all the above implicit constraints, not to mention the extra brought-in computation cost. To reduce the learning ambiguity, we propose to localize the formulation of the constraint pairs to maintain the decision boundary staying global optimal. That is, as for the intra class learning problem (e.g., both F_i and $F_j \in$ **Pos**), we focus our learning iterations toward those constraint pairs with large feature distance. Then, as for the inter class cases (e.g., $F_i \in$ **Pos** and $F_j \in$ **Neg**), our learning procedure is biasing to those constraint pairs with small feature distance. Meanwhile, the formation of constraint pairs are all unique bijections with minimum global feature distance (i.e., the feature distance sum of all selected constraint pairs) to ensure the generalization ability of our learned feature transformation.

To achieve this, we utilize binary assignment strategy to construct constraint pairs, which can be detailed as Eq. 11 (total $\min\{m, n\}$ constraint pairs for inter class problem) and Eq. 12 (total $m + n$ constraint pairs for intra class problem).

$$\min_{\mathbf{Q}} \ \Theta\left(\begin{bmatrix} ||F_1, F_1||_2, & \cdots, ||F_1, F_n||_2 \\ & \ddots & \\ ||F_m, F_1||_2, & \cdots, ||F_m, F_n||_2 \end{bmatrix}\right), \tag{11}$$

where $\Theta(\cdot)$ is the column-wise selective function which performs the binary assignment $\mathbf{Q} \in \mathbb{R}^{1 \times \min\{p,n\}}$ to formulate our inter class constraint pairs. Also, the formulation of the intra class constraint pairs can be viewed in Eq. 12.

$$\max_{\mathbf{Q}} \ \Theta\left(\begin{bmatrix} ||F_1, F_1||_2, & \cdots, ||F_1, F_{m|n}||_2 \\ & \ddots & \\ ||F_{m|n}, F_1||_2, & \cdots, ||F_{m|n}, F_{m|n}||_2 \end{bmatrix}\right). \tag{12}$$

Actually, the above optimization can be efficiently solved by Hungarian algorithm [31] in polynomial time, and thus the constraint pairs are explicitly indicated by the binary assignment result $\mathbf{Q}$.

C. Mathematical Solver

To solve Eq. 10, we propose to use the Bregman projections [46], which project the current solution onto single constraint, to simultaneously satisfy those explicit constraint pairs and maintain closeness between $\mathbf{A}$ and $\mathbf{I}$. And the iteration steps of the Bregman projection toward single distance constraint (e.g., (F_i, F_j)) can be represented as follows:

$$\mathbf{A} \leftarrow \mathbf{A} + \mu \mathbf{A}(F_i - F_j)(F_i - F_j)^T \mathbf{A}, \tag{13}$$

where μ controls the direction and strength of the Bregman projections, which can be detailed as follows:

$$\mu_{i,j} = \frac{\kappa_{i,j}\alpha}{1 - \kappa_{i,j}\alpha \times w}, \tag{14}$$

$$\kappa_{i,j} = \begin{cases} 1, & if \quad i \in \textbf{Pos} \ and \ j \in \textbf{Neg} \\ -1, & if \ (i,j) \in \textbf{Pos} \ or \ (i,j) \in \textbf{Neg} \end{cases}. \tag{15}$$

Here, κ is the indicator parameter (Eq. 15), α and w can be iteratively updated according to our bMRF saliency assumptions via the following steps:

$$\alpha = \min(\lambda_{i,j}, \frac{\kappa_{i,j}\gamma}{\gamma + 1}(\frac{1}{w} - \frac{1}{\xi_{i,j}})) \tag{16}$$

$$w = (F_i - F_j)^T \mathbf{A}(F_i - F_j), \tag{17}$$

where w (Eq. 17) represents the projection residual under the current solution $\mathbf{A}$. Also, the slack variable ξ and the threshold λ, which is initialized to 0, can be updated via Eq. 18.

$$\lambda_{i,j} \leftarrow \lambda_{i,j} - \alpha, \quad \xi_{i,j} \leftarrow \frac{\gamma \xi_{i,j}}{\gamma + \kappa_{i,j}\alpha\xi_{i,j}}. \tag{18}$$

Specially, parameter γ is a predefined parameter to control the trade-off between satisfying the constraints and minimizing $\min_{\mathbf{A} \succeq 0} ||\mathbf{A}, \mathbf{I}||_2$ (Eq. 10), and it will be further discussed in our experiment section. In our implementation, the above learning iterations will stop when either the total error reaches the objective (10^{-2}) or the iteration times exceed the predefined allowance (i.e., $100 \times (m + n + \min\{m, n\})$).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

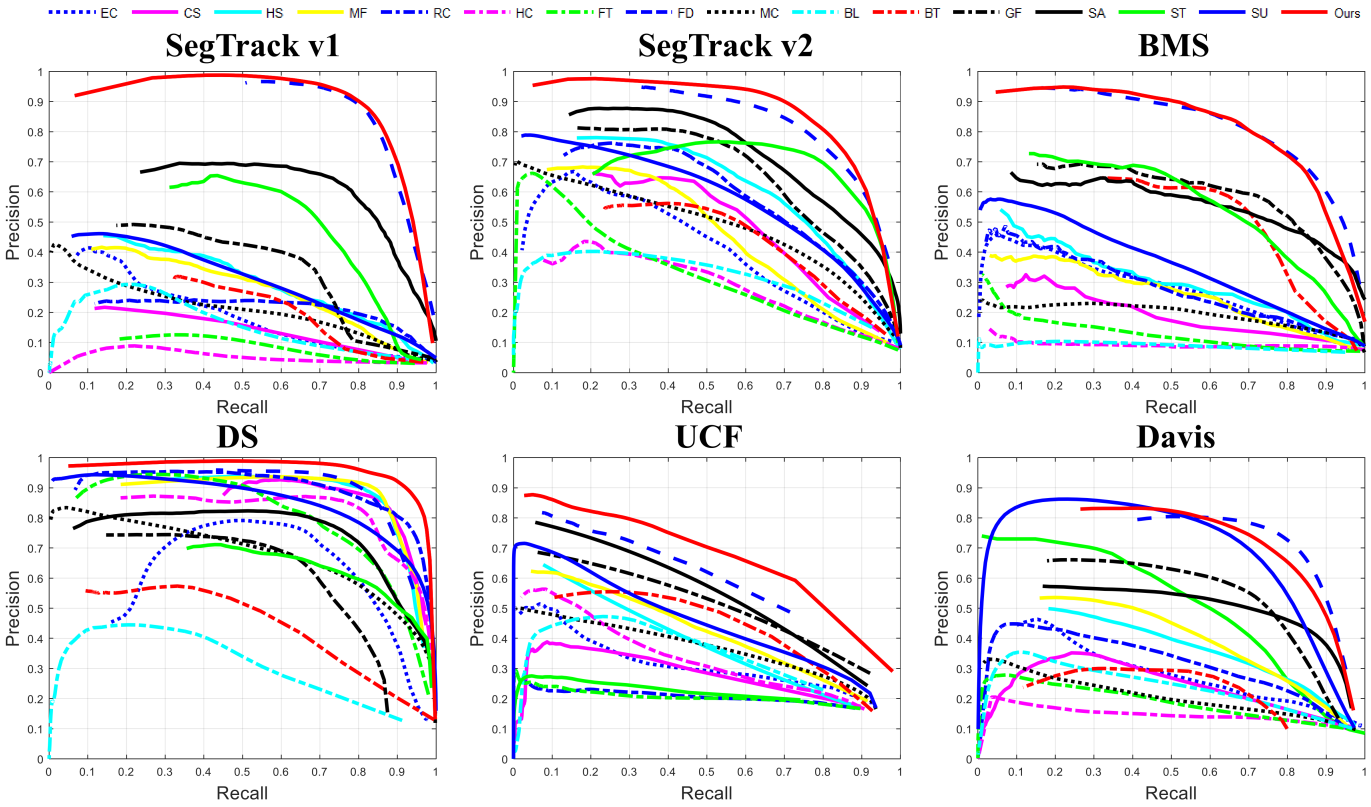

Fig. 8. Quantitative comparisons between our methods and 17 state-of-the-art methods over SegTrack v1 [32], SegTrack v2 [33], BMS [34], DS [35], UCF [36] and Davis2016(480p) [37] dataset (almost 250 video sequences). Those state-of-the-art methods include: FD17 [7], SA15 [9], GF15 [10], BT16 [22], ST14 [38], BL14 [19], MC15 [13], SU14 [39], CS13 [40], HS13 [41], MF13 [42], SB14 [18], MO13 [21], EC10 [43], RC11 [44], HC11 [44], and FT09 [45].

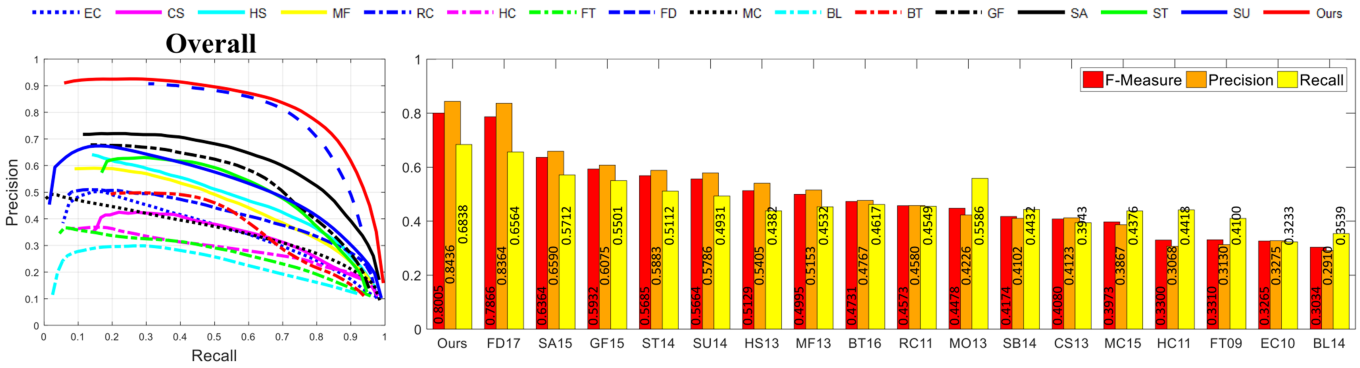

Fig. 9. Averaged quantitative comparisons between our methods and 17 state-of-the-art methods, i.e., the left part is the overall PR curve (over 6 adopted dataset) and the right part is the corresponding Precision rate and Recall rate with fixed thresholds according to the largest F-measure. Those state-of-the-art methods include: FD17 [7], SA15 [9], GF15 [10], BT16 [22], ST14 [38], BL14 [19], MC15 [13], SU14 [39], CS13 [40], HS13 [41], MF13 [42], SB14 [18], MO13 [21], EC10 [43], RC11 [44], HC11 [44], and FT09 [45].

D. Learned Video Saliency

So far, for each video frame batch, we have independently learned one non-linear feature transformation **A**, which can automatically enlarge the inter class' feature distance while maintain compactness of intra class cases.

Therefore, with the learned feature transformation **A**, we should assign large saliency value to those undetermined regions **Unk** if it exhibits large transformed feature distance toward the **Neg**. That is, we resort instances of **Neg** as the contrast basement to compute the learned video saliency (Sal_i) as follows:

$$Sal_i = \frac{1}{Z} \sum_{j \in \mathbf{Neg}} (F_i - F_j)^T \mathbf{A} (F_i - F_j), \tag{19}$$

where Z represents the normalization factor. Although Eq. 19 can correctly assign saliency value to those undetermined regions **Unk** in most cases, it may produce incorrect saliency assignments if **Unk** is extremely distinctive to both **Pos** and **Neg** in the current video frame batch. To conquer this limitation, we propose to integrate multiple feature transformations, which are complementary learned from consecutive frame

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

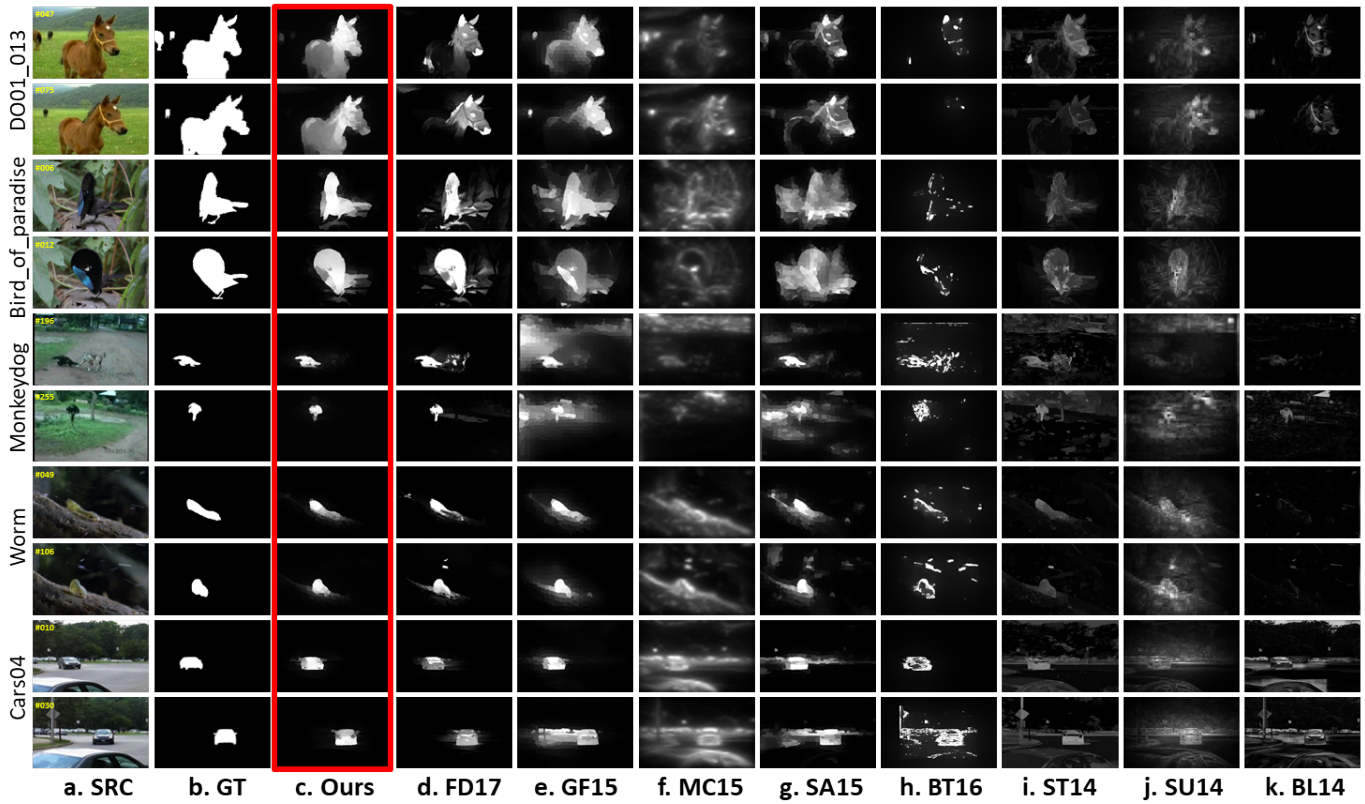

Fig. 10. Qualitative comparisons over SegTrack v1 [32], SegTrack v2 [33], BMS [34], and DS [35] datasets. **SRC** denotes the source input video frames, **GT** shows the ground truth, **Ours** demonstrates the results obtained by our method (highlighted with red rectangle), and column **d-k** demonstrate some state-of-the-art methods, including: FD17 [7], GF15 [10], MC15 [13], SA15 [9], BT16 [22], ST14 [38], SU14 [39] and BL14 [19].

batches, into current saliency computation (the t-th frame batch), see Eq. 20.

$$Sal_i = \frac{1}{N} \sum_{k=t-2}^{t+2} \eta_k \sum_{j \in \mathbf{Neg}_k} (F_i - F_j)^T \mathbf{A}_k (F_i - F_j), \quad (20)$$

where we empirically set the weight parameter η as η_{t-2}=η_{t+2}=0.1, η_{t-1}=η_{t+1}=0.25, η_t=0.3. To sharpen the salient objects boundary and slightly suppress the false-alarm detections, we also conduct spatial-temporal smoothing and pixel-wise spatial-temporal smoothing scheme, which is identical to Eq. 3, and the pictorial demonstrations can be found in Fig. 4. Also, the proofs toward the performance improvement brought by introducing the above learning scheme can be found in Fig. 14.

Here we further demonstrate the intuition of our long-term metric learning toward the rejection of non-salient regions in Fig. 6. In fact, the behind rational of our novel method is to utilize bMRF to obtain trustworthy salient and non-salient semantic labels (i.e., **Pos** and **Neg**), then utilize these labels (specially those beyond scope binary assumptions) to guide the subsequent metric learning to reveal the long-term common consistency of the salient foregrounds. Based on the learned long-term saliency consistency, our method can either compress those short-term motion induced false-alarm detections (i.e., via contrast computation under the learned long-term feature transformations, i.e., #142 frame in Fig. 6) or fill the intermittent movement caused hollow effect (demonstrations in Fig. 2)

VI. EXPERIMENTS AND EVALUATIONS

A. Parameter Selection

In principle, there are a total of three parameters having influence on the performance of our method: bMRF' strength parameters β_c, β_a (Sec. IV-B) and learning balance factor γ (Eq. 16). Since our learning performance is dependent on the bMRF computed semantic labels, we first to quantitatively test the overall performance toward different combinations of β_c and β_a, and then determine the optimal choice of γ.

Parameter β_c and β_a. We have quantitatively tested the performance of these parameters to obtain an optimal choice, and the evaluation results can be found in Fig. 12a, where the tested combinations include: $\beta_c \in \{10, 15, 20\}, \beta_a \in \{1.5, 2, 2.5\} \times \beta_c$. In fact, a large choice of β_c easily affects the integrity of the detected video saliency, while a small choice tends to produce massive undetermined regions which easily lead to poor learning performance. Meanwhile, a large choice of β_a definitely reduces the amount of determined positive salient foregrounds (**Pos**) at the expense of confidence degree, and vice versa. Thus, according to the results demonstrated in Fig. 12a, we set the optimal choice of $\beta_c = 10$ and $\beta_a = 2 \times \beta_c$. **Parameter γ.** Actually, as we mentioned before, the parameter γ controls the tradeoff between satisfying the constraints and minimizing $\min \ ||\mathbf{A}, \mathbf{I}||_2$. That is, a large choice of γ easily leads the optimization process biasing toward the constraints, and vice versa. However, being observed in Fig. 12b, these quantitative results suggest multiple

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

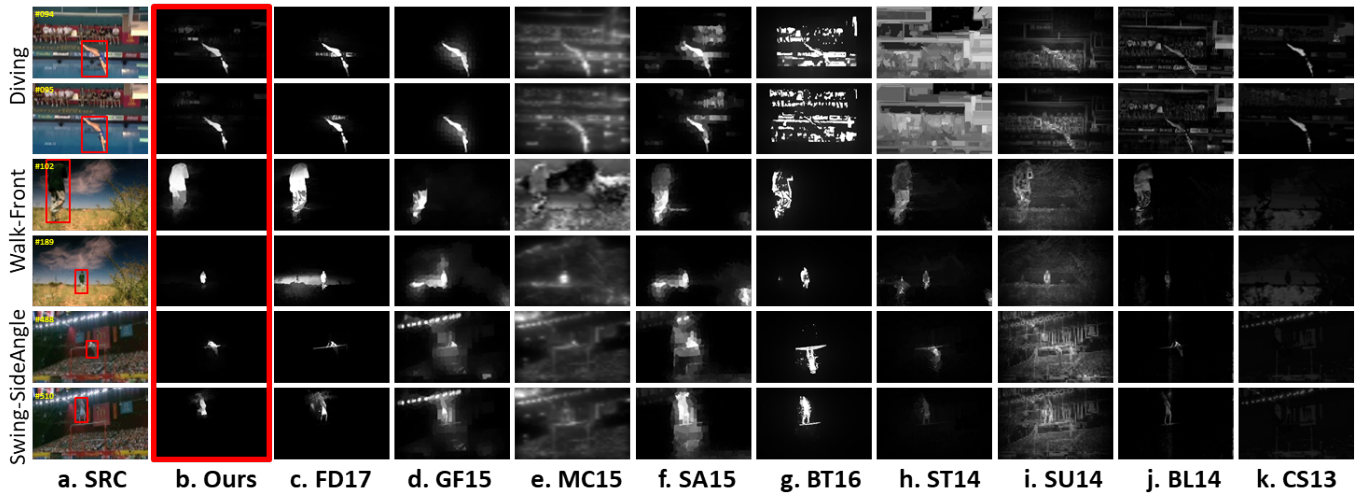

Fig. 11. Qualitative comparisons over UCF [36] dataset. **SRC** denotes the source input video frames with Ground Truth marked with red rectangle, **Ours** demonstrates the results obtained by our method (highlighted with red rectangle), and column **c-k** demonstrate some state-of-the-art methods, including: FD17 [7], GF15 [10], MC15 [13], SA15 [9], BT16 [22], ST14 [38], SU14 [39], BL14 [19] and CS13 [40].

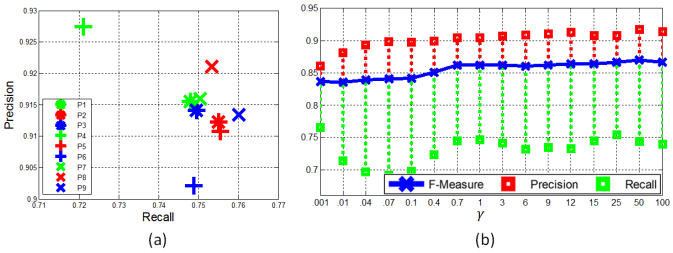

Fig. 12. (a) Quantitative evaluation toward different combination of β_a and β_c (Eq. 5), where **P1**, **P2** and **P3** respectively represent: $\beta_c \in \{15\}, \beta_a \in \{1.5, 2, 2.5\} \times \beta_c$; **P4**, **P5** and **P6** respectively represent: $\beta_c \in \{20\}, \beta_a \in \{1.5, 2, 2.5\} \times \beta_c$; **P7**, **P8** and **P9** respectively represent: $\beta_c \in \{25\}, \beta_a \in \{1.5, 2, 2.5\} \times \beta_c$, and we select P8 as our optimal choice; (b) quantitative evaluation toward different choice of γ (Eq. 16).

complementary choices, i.e., $\gamma = 3, 6, 9$. Therefore, for each video frame batch, we independently learn 3 non-linear transformations with $\gamma \in \{3, 6, 9\}$ respectively via CPU parallel computing. Thus, we further formulate the computation of learned video saliency as follows:

$$Sal_i = \frac{1}{N} \sum_{k=t-2}^{t+2} \eta_k \sum_{l=1}^{3} \sum_{j \in \mathbf{Neg}_k} (F_i - F_j)^T \mathbf{A}_k^l (F_i - F_j). \quad (21)$$

After determining the aforementioned parameters, we quantitatively evaluate the overall performance of our method by testing different combinations of the component involved in our method, and the results can be found in Fig. 14, where the **Color Saliency** exhibits the worst precision-recall curve, while the **Motion Saliency** exhibits much better performance. Meanwhile, the **Low-level Saliency**, which is fused by spatial and temporal saliency clues, is much better than pure **Color Saliency** or **Motion Saliency**. Since our newly proposed learning scheme can guarantees the spatial-temporal smoothness of the computed videos saliency, the Learned Video Saliency (i.e., **Bi-level Learning** in Fig. 14) outperforms the **Low-level Saliency** by a large margin. Also, the Learned Video Saliency (**Bi-level Learning**) can be further improved by introducing our long-term learning scheme via integrating multiple beyond scope learned feature transformations, see the **Long-term Learning** in Fig. 14. Furthermore, the spatial-temporal smoothing and pixel-wise assignment strategies (**Final**) can also boost the performance of our **Long-term Learning** scheme slightly.

B. Quantitative Evaluations

In this paper, we evaluate the performance of our method over 5 public benchmarks, including SegTrack v1 [32], SegTrack v2 [33], BMS [34], DS [35], UCF [36] and Davis2016(480p) [37] dataset. The SegTrackv1 dataset contains 6 short video sequences with fast object movements compounded with complex surroundings. The SegTrackv2 dataset contains 8 video sequences with mild-level object movements in either stationary or non-stationary scenes. The DS dataset contains 10 video sequences with slow object movements and dynamic backgrounds. The BMS dataset contains 26 diverse-length video sequences with various movements. The UCF dataset, which is guided by the human eye fixations, contains almost 150 sport related video sequences (exist almost 25% multiple salient object cases). The Davis2016(480p) dataset contains 50 video sequences (all videos only contain one salient object) with well annotated GroundTruth.

We compare our method with 17 state-of-the-art methods, including FD17 [7], SA15 [9], GF15 [10], BT16 [22], ST14 [38], BL14 [19], MC15 [13], SU14 [39], CS13 [40], HS13 [41], MF13 [42], SB14 [18], MO13 [21], EC10 [43], RC11 [44], HC11 [44], and FT09 [45]. To better verify and validate the performance of our method, we leverage the well-recognized precision-recall (PR) as evaluation indicator. Towards this end, we alternatively segment the video saliency detection results of different methods with the same threshold ($\mathbf{T} \in [0, 255]$), and the regions with saliency values larger than $\mathbf{T}$ are labeled as foreground. If the obtained foreground is consistent with the ground truth mask, it is deemed as successful detection, and the final precision-recall curves are obtained by varying $\mathbf{T}$ from 0 to 255. As the recall rate is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

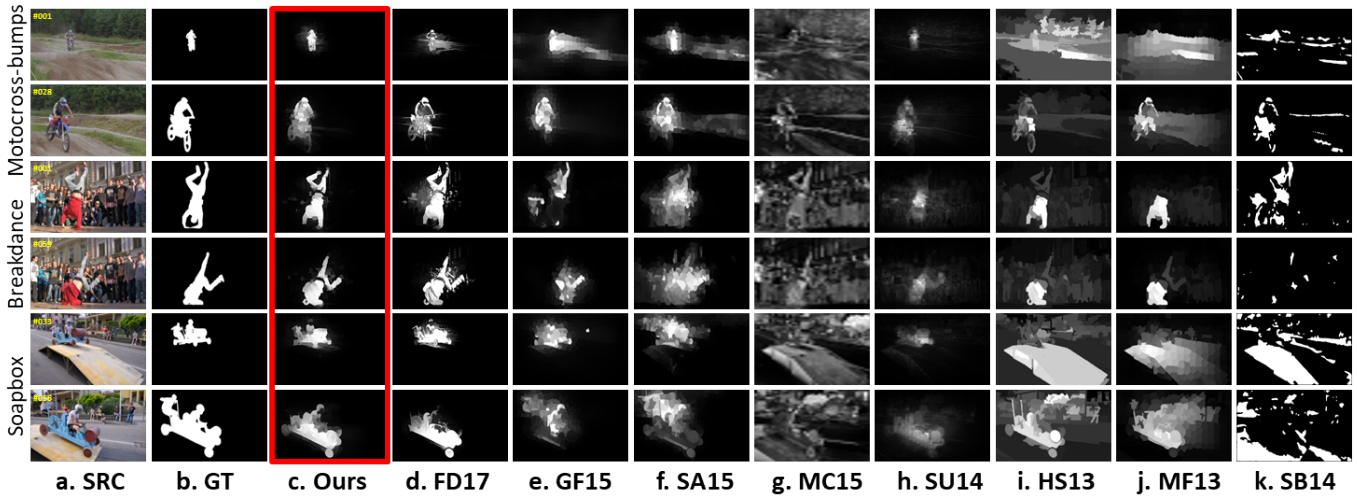

Fig. 13. Qualitative comparisons over Davis2016(480p) [37] dataset. **SRC** denotes the source input video frames, **GT** denotes the ground truth, **Ours** demonstrates the results obtained by our method (highlighted with red rectangle), and column **d-k** demonstrate some state-of-the-art methods, including: FD17 [7], GF15 [10], SA15 [9], MC15 [13], SU14 [39], HS13 [41], MF13 [42] and SB14 [18].

inversely proportional to the precision, the tendency of the trade-off between precision and recall can truly indicates the overall video saliency detection performance. We demonstrate the quantitative comparison results in Fig. 8. Since the ground truth in UCF dataset is guided by human fixations, we also reported the AUC results in Fig. 15.

As we can see the selected qualitative comparisons in Fig. 10, Fig. 11 and Fig. 13, because the conventional graph methods (i.e., GF15 and SA15) easily lead to the accumulation of false-alarm detection when pursuing the spatial-temporal smoothness of the detected video saliency, massive false-alarm detections can be easily found in their detection result, specially for those salient foreground object undergo rapid movements (e.g., the *monkeydog* sequence). Although the FD17 method adopts the low-rank guided alternative alignment strategy to avoid the accumulation problem, yet it frequently regards those short-term non-salient motions as salient foregrounds, e.g., the floats in *worm* sequence. Particularly, due to the absence of the long-term info, FD17 method easily produces hollow effect while the salient foreground object undergoes long period intermittent movement, e.g., the hollow horse in *DO01_013* sequence. As for those fusion based video saliency detection methods, e.g., ST14 and SU14, their massive false-alarm detections are mainly brought by the deficiencies of the sole fusion, which totally neglects the spatial-temporal coherency. Further, because of the absence of the temporal info, the conventional image saliency methods (e.g., HS13 and MF13) exhibit much worse detections over all adopted benchmarks. As for the results over the UCF dataset, all these compared methods exhibit low recall rate, because the human eye fixation guided ground truths are marked by a rectangle box.

Specially, the quantitative comparisons of our method toward the state-of-the-art methods over Davis2016(480p) dataset can be found in the bottom-right of Fig. 8. Since the salient objects in Davis2016 dataset are mostly dominated by plain movements, the motion clue biased video saliency detection methods (e.g., FD17 and SU14) easily obtain remarkable quantitative scores. In fact, the major advantage of our method is the newly designed learning solution to reveal long-term info to handle scenarios with complex movements (e.g., the intermittent movements in *horse* sequence in DS dataset [35]). Thus, it is reasonable for our method to exhibit equivalent performance to FD17 (motion biased method) over plain motion dominated dataset, i.e., Davis2016.

Moreover, we also leverage the average precision, recall, and F-measure indicators to demonstrate the advantages of our method. The F-measure can be computed via

$$\text{F-measure} = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \qquad (22)$$

where Precision denotes the average precision rate, Recall denotes the average recall rate, and $\beta^2 = 0.1$ to bias toward the Precision rate. It can be easily found in right-bottom of Fig. 9 that, our method apparently outperforms other state-of-the-art methods by a large margin. It also should be noted that, the detection results from both SB14 and MO13 methods are binary maps, so they have no PR curves in Fig. 8.

C. Limitations

In practice, there are total two limitations which may hinder the broad application of our method. First, our method must be applied via the off-line manner. Although we can buffer the input video frame streams to alleviate the above limitation, yet the performance trade-off still exist which deserves our future investigation. Also, Fig. 16 demonstrates failure case of our proposed method, which is mainly caused by the incorrect long-term low-level saliency estimation. Another limitation of our method is that, our method tends to be little time-consuming, i.e., our method needs about 2.6s (CUDA accelerated on a Alienware laptop with Quad Core i7-6700HQ 2.6 GHz, 16GB RAM and GTX 970m) to perform the saliency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2839523, IEEE Transactions on Multimedia

TABLE I
AVERAGE TIME COST (IN SECONDS) FOR SINGLE VIDEO FRAME OF THE STATE-OF-THE-ART METHODS. **BOLD** FONTS INDICATE THE BEST PERFORMANCE WHILE THE *italic* FONTS INDICATE THE SECOND-BEST ONES, AND THE NORMAL FONTS INDICATE THE THIRD-BEST ONES.

Method	Ours	FD17	SA15	GF15	ST14	SU14	BT16	MO13	MC15	BL14	CS13	SU14	BL14
Time Cost	*2.63*	3.93	2.56	13.3	24.31	90.3	3.47	**.291**	55.3	53.3	3.59	90.6	53.3

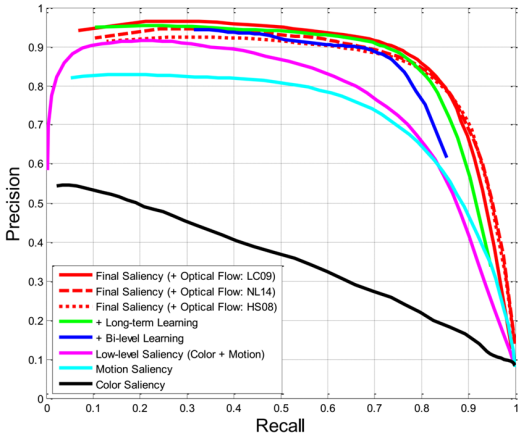

Fig. 14. Precision-recall curves of our method combining with different components, wherein **Color Saliency** represents the obtained video saliency using color saliency clues alone, **Motion Saliency** represents the obtained video saliency using motion saliency clues alone, **Low-level Saliency (Color + Motion)** represents the obtained low-level saliency, **+ Bi-level Learning** represents the learned video saliency with intra batch feature transformations only, **+ Long-term Learning** represents the learned video saliency with both intra and inter batch feature transformations (i.e., our long-term learning scheme) and **Final Saliency** represents the final video saliency detection after superpixel-wise and pixel-wise spatial-temporal smoothing scheme. We also tested our method with different optical flow choices, i.e., HS08 [47], LC09 [27] and NL14 [48]. It should be noted that all these quantitative results are evaluated over SegTrack v1 [32], SegTrack v2 [33] and BMS [34] dataset.

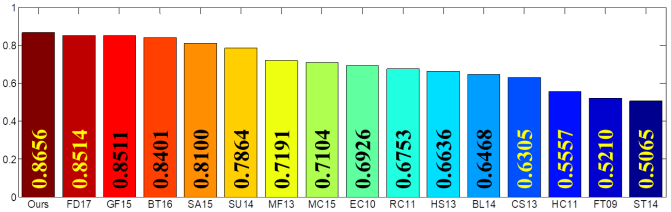

Fig. 15. Quantitative comparisons (AUC results) between our methods and the state-of-the-art methods over UCF [36] datasets. Those state-of-the-art methods include: FD17 [7], SA15 [9], GF15 [10], BT16 [22], ST14 [38], BL14 [19], MC15 [13], SU14 [39], CS13 [40], HS13 [41], MF13 [42], EC10 [43], RC11 [44], HC11 [44], and FT09 [45].

detection for single 300×300 frame. The major bottle neck of the computation relies on two dimensions: the optical flow computation (almost 1s) and the learning iterations (almost 0.9s). And the time cost comparison toward the state-of-the-art video saliency detection methods are detailed in Table. I. It should also be noted that we can alleviate time-consuming limitation via adopting deep learning based optical flow solution as paper [49], which can reduce the time cost of optical flow by half.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we have detailed a novel learning framework for video saliency detection, which takes full advantage of the long-term spatial-temporal consistency to boost the detection accuracy. Our method comprises several novel technical elements, including: (a) The newly-proposed bi-level Markov random field (bMRF) based saliency assumption can well represent the spatial-temporal consistency constraint explicitly by the binary saliency assumption; (b) Based on the bMRF guided saliency assumption, our learning solution can well utilize the intrinsic spatial-temporal smoothness to robustly compute the video saliency while avoiding the accumulation of possible false-alarm errors, and also resort to the long-term common consistency to further boost the accuracy of the detected video saliency.

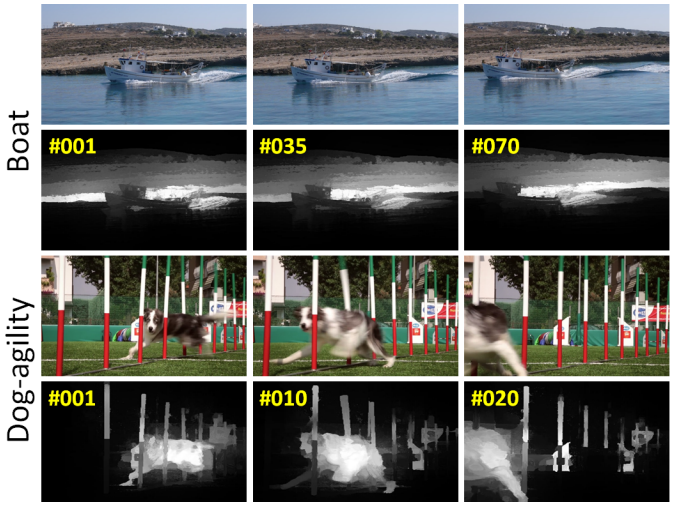

Fig. 16. Failure cases on Davis2016 [37] dataset. Since the motivation of our bMRF semantic labelling is heavily dependent on the quality of the pre-computed low-level saliency clues (i.e., motion clues and color clues), our method easily fails when both the motion saliency and the color saliency are simultaneously incorrect for all video frames.

As for our near future works, we are particularly interested in utilizing the learning-based solution to perform the motion saliency detection instead of the most time-consuming optical flow computation. Also, we are specially interested in designation of an appropriation solution to integrate the Gestalt cues [50] [51] to guide the spatial-temporal alignment and diffusion.

Acknowledgments. This research is supported in part by National Natural Science Foundation of China (No. 61190120, 61190124, 61190125, 61300067, 61672077, 6167214, 61602341, 61532002 and 61772277), Applied Basic Research Program of Qingdao (No. 161013xx) and National Science Foundation of USA (No. IIS-1715985, IIS-0949467, IIS-1047715, and IIS-1049448).

REFERENCES

[1] 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 Recognition*, vol. 48, no. 9, pp. 2885–2905, 2015.

[2] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1305–1312.

[3] H. Liang, R. Liang, and G. Sun, "Looking into saliency model via space-time visualization," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2271–2281, 2016.

[4] D. Chen and Y. Luo, "Preserving motion-tolerant contextual visual saliency for video resizing," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1616–1627, 2013.

[5] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for hevc-msp," *IEEE Transactions on Multimedia*, p. 10.1109/TMM.2017.2721544, 2017.

[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, and K. Rapantzikos, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[7] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3156–3170, 2017.

[8] V. Mahadevan and N. Vasconcelos, "Spatio-temporal saliency in dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 171–177, 2010.

[9] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

[10] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.

[11] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2303–2316, 2015.

[12] 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*, 2017.

[13] H. Kim, Y. Kim, J. Sim, and C. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, 2015.

[14] C. Aytekin, H. Possegger, T. Mauthner, S. Kiranyaz, H. Bischof, and M. Gabbouj, "Spatiotemporal saliency estimation by spectral foreground detection," *IEEE Transactions on Multimedia*, p. 10.1109/TMM.2017.2713982, 2017.

[15] S. Chaabouni, J. Benoispineau, and O. Hadar, "Deep learning for saliency prediction in natural video," 2016.

[16] W. Wang, J. Shen, and L. Shao, "Deep learning for video saliency detection," 2017.

[17] S. Varadarajan, P. Miller, and H. Zhou, "Spatial mixture of gaussians for dynamic background modelling," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2013, pp. 63–68.

[18] P. StCharles, G. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2015.

[19] Z. Gao, L. Cheong, and Y. Wang, "Block-sparse rpca for salient motion detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1975–1987, 2014.

[20] 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 *Neural Information Processing Systems*, 2009, pp. 2080–2088.

[21] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.

[22] 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 Recognition*, vol. 52, pp. 410–432, 2016.

[23] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *AAAI Conference on Artificial Intelligence*, 2013, pp. 1063–1069.

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

[25] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 628–635.

[26] E. Gastal and M. Olive, "Domain transform for edge-aware image and video processing," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 1–12, 2011.

[27] C. Liu, "Exploring new representations and applications for motion analysis," *Massachusetts Institute of Technology*, 2009.

[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels," *EPFL Technical Report*, 2010.

[29] A. Delong, A. Osokin, H. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 1–27, 2012.

[30] D. Martin, S. Fahad, F. Michael, and V. Joost, "Adaptive color attributes for real-time visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1090–1097.

[31] M. James, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[32] D. Tsai, M. Flagg, A. Nakazawa, and M. James, "Motion coherent tracking using multi-label mrf optimization," *International journal of computer vision*, vol. 100, no. 2, pp. 190–202, 2012.

[33] F. Li, T. Kim, A. Humayun, D. Tsai, and R. James, "Video segmentation by tracking many figure-ground segments," in *IEEE International Conference on Computer Vision*, 2013, pp. 2192–2199.

[34] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *European Conference on Computer Vision*, 2010, pp. 282–295.

[35] F. Ken, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 638–641.

[36] M. Stefan and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *European Conference on Computer Vision*, 2012, pp. 842–856.

[37] 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 *Computer Vision and Pattern Recognition*, 2016.

[38] F. Zhou, S. Kang, and F. Michael, "Time-mapping using space-time saliency," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3358–3365.

[39] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[40] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[41] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[42] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[43] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*, 2010, pp. 366–379.

[44] M. Cheng, G. Zhang, J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 409–416.

[45] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[46] B. Kulis and I. Dhillon, "Learning low-rank kernel matrices," in *International Conference on Machine Learning*, 2006, pp. 505–512.

[47] D. Sun, S. Roth, J. Lewis, and M. J. J. Black, "Learning optical flow," in *In European Conference on Computer Vision*, 2008, p. 83C97.

[48] D. Sun, S. Roth, and M. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, 2014.

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

[50] J. Yu, G. Xia, C. Gao, and A. Samal, "A computational model for object-based visual saliency: Spreading attention along gestalt cues," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 273–286, 2016.

[51] A. Russell, S. Mihala, R. Von, E. Niebur, and R. Etiennecummings, "A model of proto-object based saliency," *Vision Research*, vol. 94, no. 1, pp. 1–15, 2014.