Improved Robust Video Saliency Detection Based on Long-Term Spatial-Temporal Information

Chenglizhao Chen, Guotao Wang, Chong Peng, Xiaowei Zhang, and Hong Qin, *Senior Member, IEEE*

***Abstract*— This paper proposes to utilize supervised deep convolutional neural networks to take full advantage of the long-term spatial-temporal information in order to improve the video saliency detection performance. The conventional methods, which use the temporally neighbored frames solely, could easily encounter transient failure cases when the spatial-temporal saliency clues are less-trustworthy for a long period. To tackle the aforementioned limitation, we plan to identify those beyond-scope frames with trustworthy long-term saliency clues first and then align it with the current problem domain for an improved video saliency detection.**

***Index Terms*— Video saliency detection, spatial-temporal saliency consistency, low-level saliency clues, long-term information revealing.**

I. Introduction and Motivation

AS AN important pre-processing tool, salient object detection (SOD) aims to localize those most eye-attractors in the given scene [1]–[6], and its subsequent application comprises various CV problems, including tracking [7]–[9], video surveillance [10], [11], and retrieval [12]. Different to the SOD in static image using spatial information only, the newly available temporal information in video data is a critical factor to localize the salient object, making the SOD task challenging [13].

In general, almost all the conventional hand-crafted video saliency methods follow the bottom-up rationale [14]–[17], i.e., saliency clues are individually (or interactively) computed over spatial and temporal channel beforehand, and then these clues will be selectively fused as the video saliency detection result. After entering the deep learning era [18], the conventional phase-wise fashion has been significantly outperformed by the high efficiency and automatical end-to-end full convolutional network based methods [19]–[21]. Although much improvements have been made, the current main stream methods are still encountering failure cases, which we believe it is mainly caused by the "limited sensing scope". That is, due to the hard-ware limitation, the state-of-the-art methods can only use temporally neighbored several frames to conduct their saliency predictions, and we name such implementation as "short-term manner". However, the problem is that when the spatial information contradicts with the temporal information short-termly, either the automatically learned short-term deep model or the hand-crafted fusion scheme may become confusing, producing failure cases eventually. Moreover, the failure case frequency of such short-term fashion is positively related to the spatial-temporal contradiction level, limiting the broad application of video saliency in real scenes, e.g., the cases with long-period intermittent movement induced motion absence. Further, such short-term manner also contradicts with the real human vision system (HVS), e.g., the HVS may continue to focus on a specific object even in the case that the gazed object is temporally not so salient.

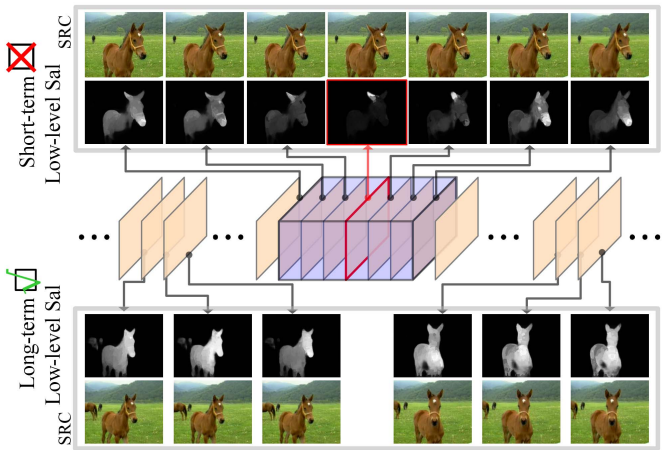

Fig. 1. The conventional "short-term" manner may be incapable to perform correct detection due to the unpredictable nature of movements, yet the beyond-scope long-term information may potentially valuable to benefit the current detection.

Thus, all the mentioned above motivates us to investigate a long-term manner, utilizing the beyond scope spatial-temporal information to facilitate the current video saliency detection. We demonstrate the methodology difference between the short-term manner and our novel long-term manner in Fig. 1. To achieve it, we propose to exploit the long-term spatial-temporal information (LSTI) towards more robust detection

Manuscript received December 5, 2018; revised June 1, 2019; accepted July 30, 2019. Date of publication August 23, 2019; date of current version November 4, 2019. This research was supported in part by the National Natural Science Foundation of China (No. 61802215 and No. 61806106), in part by the Natural Science Foundation of Shandong Province (No. ZR2019BF011 and ZR2019QF009), and in part by the National Science Foundation of USA (No. IIS-1715985 and IIS1812606). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kalpana Seshadrinathan. *(Chenglizhao Chen and Guotao Wang contributed equally to this work.) (Corresponding authors: Chong Peng; Xiaowei Zhang.)*

C. Chen, G. Wang, C. Peng, and X. Zhang are with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, China (e-mail: cclz123@163.com; qduwgt@163.com; pchong1991@163.com; xiaowei19870119@sina.com).

H. Qin is with the Computer Science Department, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: qin@cs.stonybrook.edu).

Digital Object Identifier 10.1109/TIP.2019.2934350

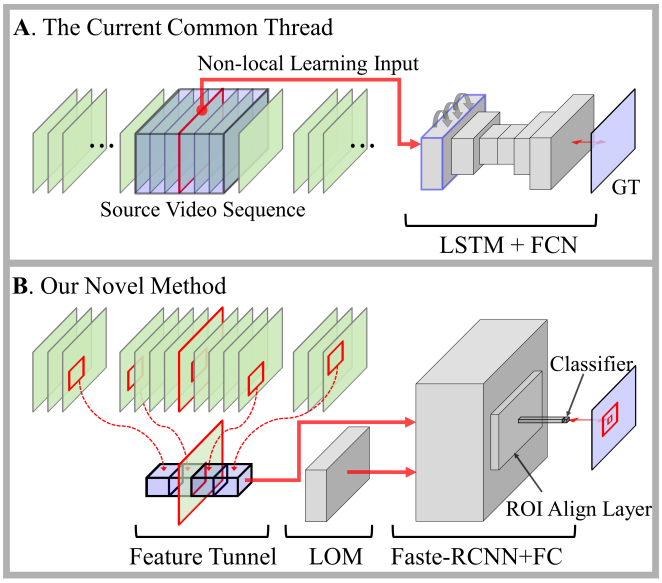

Fig. 2. Demonstration of the differences between the conventional methods based on the LSTM network and our new method.

with improved accuracy. Our LSTI is initially formulated from the beyond-scope "high-quality" low-level saliency clues (Sec. III-A), whose qualities are measured by the newly designed fast quality assessment scheme (FQA, Sec. III-B). We design a series of measures to continue to improve the accuracy of LSTI, so as to ensure the effectiveness of LSTI in the current problem domain. Meanwhile, due to the existence of strong semantic coherency between the current frame and those beyond-scope frames, we propose to utilize the non-local spatial-temporal similarity to align the LSTI with the current problem domain (Sec. III-D). Directly benefiting from the beyond-scope LSTI, those ill-detections may be easily discriminated from its non-salient nearby surroundings. By using our newly designed network (see Fig. 2(B)), which can automatically complement the current frame with the newly introduced long-term information (i.e., LSTI), we can achieve high-quality video saliency while avoiding the above-mentioned short-term induced defects (Sec. IV-C). The salient contributions of this paper are twofold:

- We propose a novel method to integrate the long-term spatial-temporal information (LSTI) into the current video saliency detection framework to cope with the obstinate short-term induced hollow effects and false-alarm cases.
- We design a new and efficient deep network, which can automatically complement the LSTI with the current local spatial-temporal information, towards automatic and robust high-performance object detection with improved accuracy in video saliency.

II. Related Work

From the early hand-crafted spatial-temporal low-level saliency clues guided methods [22] to the current state-of-the-art deep learning solutions [20], [21], [23], the core rationale of video saliency estimation remains unchange, i.e., the video saliency estimation should be able to strike the optimal complementary state between the spatial info and the temporal info.

To reveal the saliency clues from the temporal scale, Haejong and Peyman [24] proposed to compute the contrast based saliency in a predefined spatial-temporal surroundings. Similarly, Fang *et al.* [25] proposed to sense the motion info from the spatial-temporal perspective via measuring the short-term temporal residuals within consecutive video frames. Meanwhile, the optical flow based trajectories can also be utilized to robust the motion sensing, achieving remarkable video object segmentation result [26]. Following the bottom-up scheme, Shen *et al.* [27] proposed to select a small group of most representative object trajectories first, and then these trajectories were iteratively merged into fragments and clusters to guide a robust video object segmentation. Similarly, Wang *et al.* [28] adopted a semi-supervised manner (by using a newly designed probabilistic model and clustering method) to further robust the computed trajectory, achieving remarkable performance improvements.

Rather than focusing on the designation of the motion revealing strategy, the optical flow guided contrast computation leading the performance for a long period. Zhou *et al.* [29] proposed to compute multi-scale contrast computation over two consecutive frames sensed optical flow result to reveal motion saliency. Once the motion saliency has been computed, it can be fused with color saliency to robust the detections [14]. In [25], the computed motion saliency was adaptively merged with the color saliency by using entropy-based uncertainty weights. Similarly, more intuitive fusion methods were proposed to integrate the motion saliency with the color saliency, e.g., by performing either multiplicative, additive, or maximum based fusions [29]–[31]. However, all these fusion solutions are limited when both the motion saliency and color saliency are simultaneously untrustworthy, producing massive failure detections.

To strike a better fusion result, the concept of short-term was proposed in recent years, and its core rationale is to use the spatial-temporal consistency between consecutive video frames to robust the fusion result. In order to incorporate the spatial consistency into the fusion procedure, the graph/CRF energy minimization [13], [32], the low-rank guided batch-wise alignments [15], the MRF guide metric learning [33], and the non-local random forest regressor [34] were proposed and achieved remarkable performance improvements. However, due to the varying nature of both the video scenes and the movement patterns, the above mentioned hand-crafted solutions seem to encounter the performance bottle neck.

In fact, by using the deep learning frameworks [35], [36], the video saliency detection performance can be improved significantly. Le and Sugimoto [37] proposed to use 3D convolution to automatically extract the temporal high discriminative features. However, because its spatial deep features are computed by RCNN, the learned saliency model is consisted by a large mount of tunable parameters burdening the computation. To solve this problem, Song *et al.* [38] proposed to utilize the Pyramid Dilated Convolution with ConvLSTM to simultaneously sense both the spatial info and the temporal info, which outperformed Le's work by a large margin while

running in real-time speed. Similarly from the spatial-temporal perspective, Jian *et al.* [39] proposed a two-stream fully convolutional neural network to enhance the saliency consistency. Li *et al.* [21] proposed to utilize the newly revised LSTM network to automatically integrate the non-local spatial info into the short-term temporal info, and its processing speed is also extremely fast.

Specially, although the above mentioned Full Convolution Network based methods [20] have achieved significant performance improvement, it still seem to encounter 2 major limitations, i.e.,, the shortage of well annotated videos for model training and the boundary blur of the saliency prediction. Actually, the shortage of training dataset problem can be alleviated by either adopting additional training dataset [39] or using synthetic data [20], e.g., the most recent [40] has utilized the human eye fixation [41] to facilitate the video saliency deep model training, archiving the leading performance. Also the boundary blur problem can be effectively alleviated by using the CRF refinement [35], [37].

Although much achievements have been made by the above mentioned efforts, the saliency revealing scope of the current state-of-the-art methods are still limited within the short-term scope, and massive failure detections can be easily found toward those scenarios with complex backgrounds while the temporal info undergoes long-period untrustworthy state. Thus, Chen *et al.* [33] proposed to use the metric learning solution to respectively learn the short-term common consistency of the salient foregrounds and then pool all these learned saliency model to represent the long-term info to improve the detection performance. However, since the adopted metric learning itself is heavily dependent on the intermediate assumption of the non-salient near surroundings, the newly revealed long-term info frequently becomes invalid due to the varying nature of the video scenes.

Inspired by the established inter-image non-local correspondences for image saliency detection [42], this paper propose to use the inter-frame semantic coherency to integrate the long-term spatial-temporal info (LSTI) into the current problem domain while utilizing the deep learning framework to robust its usage. By doing this, we can easily solve the above mentioned limitations and further improve the detection performance.

III. The Long-term Spatial-Temporal Info Revealing and Aligning

Since we propose to reveal our long-term spatial-temporal info (LSTI) from the pre-computed low-level saliency, we will introduce the detailed low-level saliency computation here.

A. Low-Level Saliency Computation

Given an input video sequence, we utilize SLIC [43] method to over segment each video frame into mid-level superpixels. And then, we follow the work [15] to compute the motion saliency clues (**MS**) via performing superpixel based non-local contrast computation over the Optical Flow [44] sensed temporal info. Meanwhile, we directly utilize the pre-learned image saliency model, i.e., the DSS method [36], to represent the corresponding color saliency (**CS**). Thus, we can formulate the low-level saliency (**LS** $\in \mathbb{R}^{1\times NS}$, and NS represents the superpixel number) via multiplicative based fusion as Eq. 1.

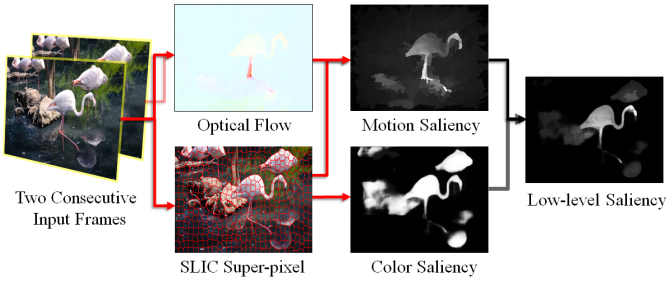

Fig. 3. The main procedure of low-level computation.

$$\mathbf{LS} = \xi(\mathbf{MS}) \odot \xi(\mathbf{CS}), \tag{1}$$

where ξ represents the $min\text{-}max$ normalization and the $\odot$ denotes the element-wise Hadamard product. We also demonstrate the pipeline of the low-level saliency computation in Fig. 3.

Benefiting from the multiplicative based fusion mechanism, those non-salient backgrounds with inconsistent **MS** and **CS** can be easily compressed. Thus, in most cases, the low-level saliency **LS** can well handle those camera view angle change induced false-alarm detections. However, due to the multiplicative nature of Eq. 1, the **LS** itself frequently encounters the hollow effects toward those salient object undergoing intermittent movements. Meanwhile, because the overall quality of the computed **LS** frequently varies with both the scene complexity degree and the current movement pattern, only those beyond scope "high quality" **LS** are potentially valuable to be regarded as the long-term spatial-temporal info (LSTI) to benefit our subsequent detection in the current frame. Therefore, we propose to utilize our newly designed fast quality assessment scheme (Sec. III-B) to filter those "low quality" **LS** before revealing the LSTI from it.

B. Fast Quality Assessment (FQA)

Since our low-level saliency computation is based on the multiplicative fusion scheme, those high quality low-level saliency **LS** must correlate to high quality spatial-temporal saliency assumptions simultaneously. Thus, the quality degree of the **LS** (Eq. 1) is intuitively positively correlated to the consistency degree between the **MS** and the **CS**. And we propose to utilize Eq. 2 to measure such consistency degree efficiently.

$$Cons = ||\max\{\mathrm{T} - \mathrm{abs}(\Lambda(\mathbf{CS}) - \Lambda(\mathbf{MS})), 0\}||_0/\vartheta, \tag{2}$$

where $Cons \in [0, 1]$ denotes the consistency degree between the **MS** and the **CS**, and T is a predefined hard threshold. Specially, $\Lambda(\cdot)$ in Eq. 2 denotes the dynamic binarization function, $\vartheta \leftarrow \max\{||\Lambda(\mathbf{CS})||_0, ||\Lambda(\mathbf{MS})||_0\}$ denotes the number of the non-zero elements of the video frame after using $\Lambda(\cdot)$ function. To this end we can utilize Eq. 2 to measure the quality degree of the given **LS**, however, we found exceptions that the real quality of **LS** may contradict with the $Cons$

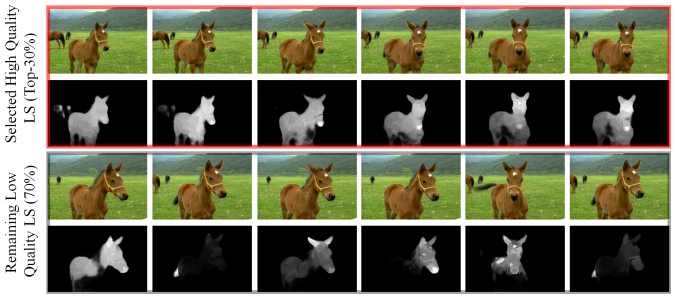

Fig. 4. Pictorial demonstrations toward our fast quality assessment scheme selected high-quality low-level saliency.

degree, which is mainly induced by the fact that our *Cons* degree may bias toward to those frames with small ϑ value. To solve this problem, we sort all *Cons* by descending order firstly, and then we select the Top-50% as the high-quality subgroup. Then, for this high-quality subgroup, we utilize the ϑ value to perform the second round ranking by descending order and only the Top-30% will be potentially used as LSTI. We demonstrate the selected high-quality **LS** in Fig. 4.

C. The LSTI Formulation

Given a video frame, the corresponding long-term spatial-temporal info (LSTI) can be revealed from those high-quality "beyond scope frames" (i.e., the frame interval > 16). Then, these newly revealed LSTI will be aligned into the current problem domain to robust the detection performance. Here we adopt the SIFT-Flow [45] method to perform the inter-frame pixel-wise alignments as Eq. 3.

$$[\mathbf{Q}_{t,\upsilon}, \mathbf{E}_{t,\upsilon}] = SF(\mathbf{I}_t, \mathbf{I}_\upsilon), \quad s.t. \quad abs(t-\upsilon) > 16, \tag{3}$$

where $SF(\cdot)$ represents the SIFT-Flow algorithm, $\mathbf{I}_t$ denotes the t-th current frame, $\mathbf{I}_\upsilon$ denotes one of the FQA scheme (Sec. III-B) selected long-term frames, $\mathbf{Q} \in \mathbb{R}^{W\times H\times 2}$ represents the inter-image pixel-wise alignment info, $\mathbf{E} \in \mathbb{R}^{W\times H\times 1}$ represents the corresponding alignment error, W and H denote the width and height of the video frame respectively.

By using Eq. 3, we can respectively obtain one pixel-wise alignment matrix $\mathbf{Q}$ for each frame pair, i.e., $\{\mathbf{I}_t, \mathbf{I}_\upsilon\}$. Actually, the SIFT-Flow method provided inter-frame correspondences are reliable in general due to the intrinsic semantic coherency within the video data. However, due to the varying nature of the salient foregrounds, the SIFT-Flow guided correspondences may become untrustworthy when the salient foregrounds undergo fast movement with drastic scale variation simultaneously. To alleviate this problem while maintaining the efficiency, we propose to use Eq. 4 to measure the non-local saliency-aware alignment quality and then the Top-m bests are selected as the revealed LSTI.

$$Alq_{t,\upsilon} = ||\mathbf{E}_{t,\upsilon} \odot G(\mathbf{LS}_\upsilon)||_1 / ||G(\mathbf{LS}_\upsilon)||_1, \tag{4}$$

where $Alq_{t,\upsilon}$ represents the frame-level alignment quality between the current t-frame and the υ-th beyond scope frame, the alignment error matrix $\mathbf{E}$ can be computed by Eq. 3, and the function $G(\cdot)$ converts the superpixel-wise **LS** into $W \times H$ pixel-wise matrix. Meanwhile, all video frames are down-sampled to 150×150 before performing the SIFT-Flow alignment to alleviate the computation burden. So far, we have aligned those valuable LSTI to the current problem domain by using Eq. 3 and Eq. 4, and then the superpixel-wise non-local alignments can be assembled accordingly.

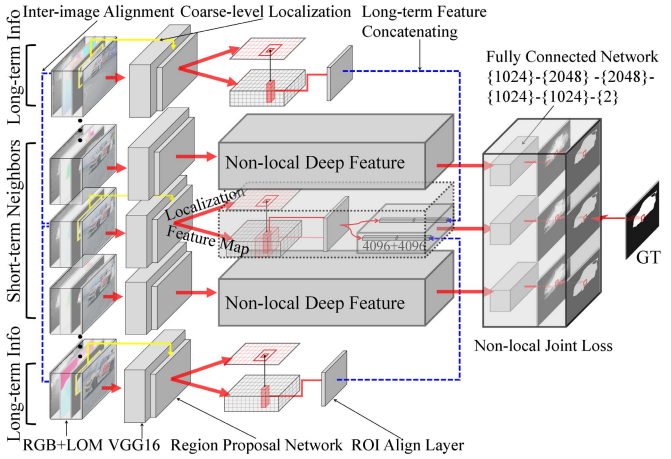

Fig. 5. The architecture of our newly designed network. Our network can simultaneously take full advantage of the LSTI usage while adopting the spatial-temporal smoothness constraint to produce high performance video saliency detection.

D. Non-Local LSTI Alignment

By using Eq. 3 and Eq. 4, we can explicitly formulate the proposed LSTI into amount of "instance pairs" and then regard it as the training dataset to train the proposed video saliency detection model. However, in our observation (see quantitative details in TABLE. III), we notice that almost 13% previously formulated LSTI are "incorrect", e.g., salient regions aligned with non-salient backgrounds, which is mainly caused by the incapability of the SIFT-Flow algorithm to handle both the scale variation and the large displacement. Thus, we propose to adopt the coarse-to-fine solution to alleviate this problem. That is, instead of directly performing the SIFT-Flow alignments to formulate our LSTI, we propose to utilize the Region Proposal Network (RPN) [46] provided object-level prior to coarsely locate the salient foregrounds beforehand (see Coarse-level Localization in Fig. 5). For each video frame, we select one RPN object proposals with largest IOU rate toward the pre-computed low-level saliency info as the coarsely located non-local region (NR). Then, by performing fine-level alignments (i.e., SIFT-Flow alignments) between two aligned NRs (see Eq. 5), we can effectively alleviate the scale-variation/large-displacement induced miss-alignments.

$$[\mathbf{Q}_{t,\upsilon}, \mathbf{E}_{t,\upsilon}] = SF(\mathbf{NR}_t, \mathbf{NR}_\upsilon), \quad s.t. \; abs(t-\upsilon) > 16. \tag{5}$$

Meanwhile, in order to avoid using the incorrect LSTI as training data, we further filter those already aligned non-local regions with large $\mathbf{E}$ via using dynamic thresholding $\mathbf{T}_E$:

$$\mathbf{Q}(x)_{t,\upsilon} \leftarrow \mathbf{Q}(x)_{t,t} \; if \; \mathbf{E}(x)_{t,\upsilon} > \mathbf{T}_E, \tag{6}$$

where x denotes the corresponding elements of the given alignments matrix, $\upsilon 1, \upsilon 2$ respectively denotes the Top-2 selected

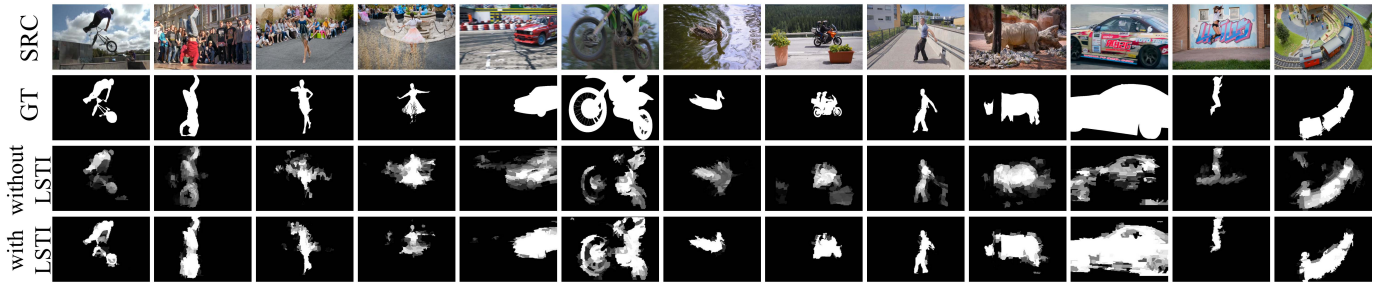

Fig. 6. Qualitative demonstration toward the effectiveness of our LSTI strategy. Since our newly integrated long-term info can emphasis on the common saliency consistency from both the spatial and the temporal perspective, those intermittent movements (mainly from the non-salient backgrounds with plain camera movement) induced false-alarms can be correctly suppressed while highlighting the salient objects.

Algorithm 1 Main Steps of LSTI Revealing and Aligning

For each video frame
1. Use Eq. 1 to compute the low-level saliency $\mathbf{LS}_i$;
2. Use the FQA scheme (Sec. III-B) to locate key frames;
3. Formulate LSTI from Step. **2** determined key frames using Eq. 4;
4. Select one RPN box with largest **LS** IOU rate as the key object proposal, Sec. III-D;
5. Align the current object proposal with the LSTI key object proposals, Eq. 5.
End For

frames by Eq. 4. We empirically assign $\mathbf{T}_E$ as $\frac{1.2}{NS} \times \min\{Alq_{t,v1}, Alg_{t,v2}\}$, and NS denotes the superpixel number. So far, we can improve the LSTI error rate from the previous 13% (Frame-level Align) to the current 3.5% (Non-local Align with EM), see quantitative proofs in TABLE. III. We also provide the a pseudo code to describe the main steps of our LSTI Revealing and Aligning as Algorithm 1.

To this end, the key proposals are already selected solely considering the pre-estimated low-level saliency clues, i.e., the **LS** (Eq. 1). Although such implementation is efficient and effective in most cases, the **LS** itself may not always robust to guide the key proposal selection, specially for the non key frames. Moreover, the video spatial-temporal consistency degree also affects the robustness of our non-local LSTI alignment accuracy, thus the revealed LSTI may become less-trustworthy for videos with weak spatial-temporal consistency, e.g., video sequences with intermittent target disappearance. Inspired by previous works [41], [47], we can resort the saliency fixation, as a complementary clue for the **LS**, to robust our non-local LSTI alignment. That is, when we conduct non-local LSTI alignment (mentioned in Sec. III-D) for sequences with weak spatial-temporal consistency, it is advisable to re-rank the IOU order (i.e., **LS** & RPN box) by further considering the fixation clue (predicted by any off-the-shelf method) as an alternate localization clue.

IV. Our Novel Deep Saliency Model

A. Intuitive Deep Features

Being noticed that the moving objects can easily attract the attention of human vision system even in scenario with complex backgrounds, the designed saliency model should bias toward the temporal info when considering the LSTI integrated deep features. Thus, we formulate the input data layer to receive 6 channel inputs (RGB+LOM, see Fig. 5), including the RGB info, the low-level saliency (L), the optical flow gradient (O), and the motion saliency (M). Since the spatial common consistency of the salient foregrounds can be revealed from the RGB info, the off-the-shelf LOM is able to effectively shrink the problem domain while avoiding the learning ambiguity.

Supposing we have two aligned superpixels, i.e., $[SP_{t,i}, SP_{v,j}]$, we utilize the multi-scale VGG16 network to formulate its convolutional deep features to handle the scale problem. Thus, its corresponding deep features can be represented as $f_{t,i} = [f_{local}, f_{mid}] \in \mathbb{R}^{1\times(4096+4096)}$, where f_{local} and f_{mid} respectively denote the local info (represented by a bounding box containing the given superpixel tightly) and the mid-level info (with bounding box 4 times larger than the local one) of the given superpixel. Thus, the deep feature of LSTI can be coarsely represented by concatenating $f_{t,i}$ with its inter-image aligned $f_{v,j}$, and then we can feed it to the ANN network for the video saliency classification. We denote the above intuitive model as "LSTI", and its qualitative demonstrations can be found in Fig. 6 and quantitative results can be found in Fig. 9.

B. Deep Feature Computation Acceleration

Given a video frame, it is time consuming (almost 6s for a single 300×300 video frame) to directly use the above mentioned strategy to conduct LSTI integrated deep feature computation for each superpixel, not to mention the multi-scale feature computation induced additional computation costs. To solve this problem, we adopt the revised VGG16 + ROIPooling [48] framework to obtain the deep feature, and we list the main modifications as following: First, we replace the final ROIPool layer to the ROIAlign layer [49] to robust the scale problem while compressing the miss-aligned cases (almost 3.5% in our training dataset); Second, we utilize the RPN network to integrate the object-level prior into our LSTI alignment steps, which can effectively reduce the problem size by half; Third, for each superpixel, we directly feed the bi-scale object proposals to the ROIAlign layer to obtain the two 4096-dimensional deep features to represent the high-level semantics info. By adopting the above mentioned steps, we can accelerate the deep feature computation by almost 10 times.

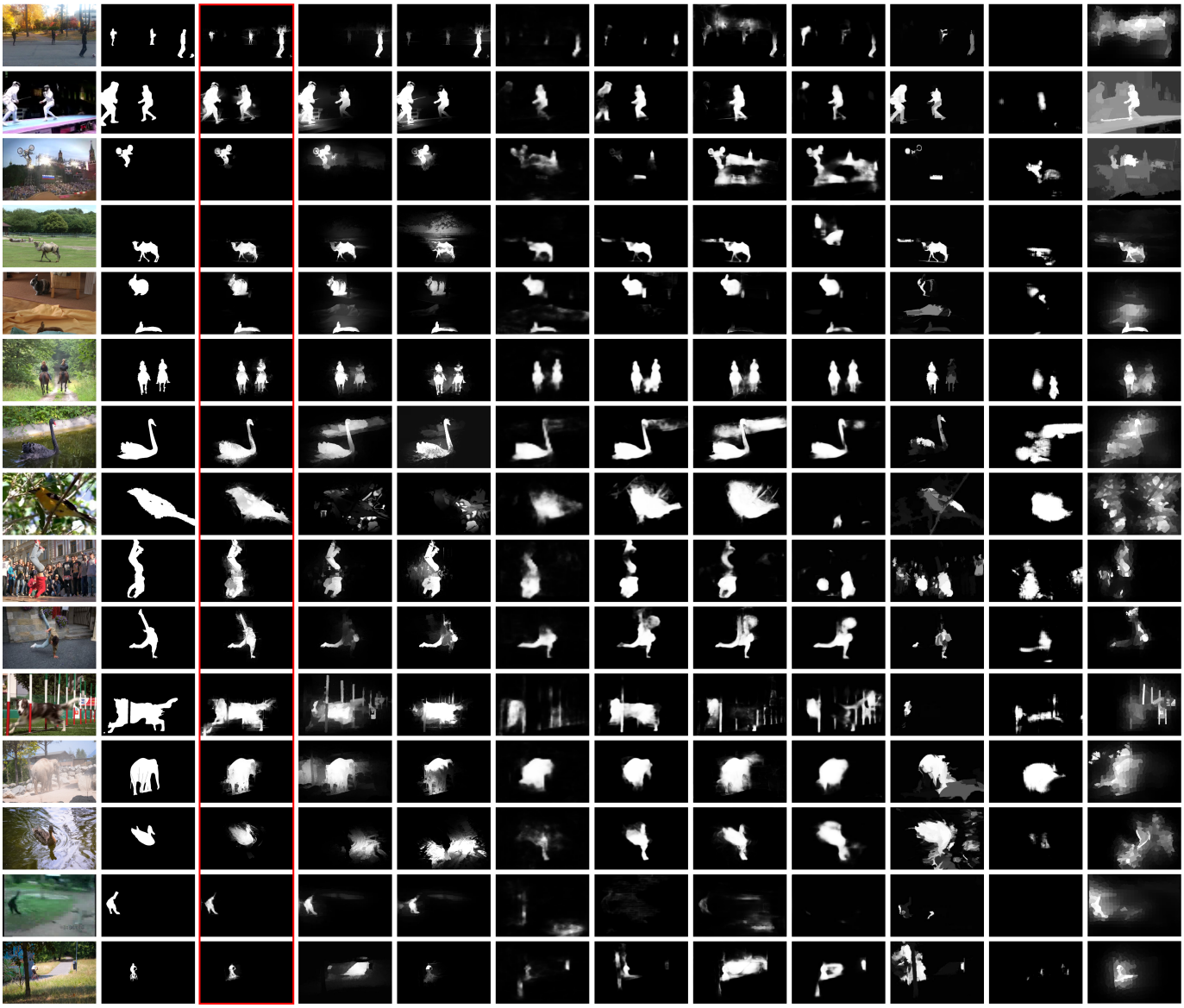

Fig. 7. Qualitative comparisons between our method and the state-of-the-art methods, where (a) denotes the source input video frame, and the human annotated binary ground truths (GT) are demonstrated in (b), (c) demonstrates the results obtained by our method (highlighted with red rectangle), and some state-of-the-art methods, including RADF18 [53], DLVSD18 [20], FL18 [33], FD17 [15], DSS17 [36], RFCN16 [55], DHS16 [54], MDF15 [35] and GF15 [13].

C. LSTI Guided Spatial-Temporal Deep Saliency Learning

To robust the usage of mentioned above LSTI, we propose to align multiple (m) beyond scope long-term info to the current video frame. Thus, we can represent the $SP_{t,i}$' deep feature of the LSTI F as $(f_{t,i}, f_{v,j})$, $j = 1, 2, ..., m$, where m denotes the maximum inter-image alignment number of the long-term info.

Therefore, the non-local joint loss function of the LSTI can be formulated by Eq. 7.

$$L_{non} = -\sum_{i=1}^{NT}\sum_{j=1}^{m}\{Z_i \log Pr(S_{i,j} = 1|F_{i,j}, \Phi) + (1 - Z_i) \log Pr(S_{i,j} = 0|F_{i,j}, \Phi)\} \cdot Alq_{i,j}, \quad (7)$$

where Z is the corresponding groundtruth label, Pr denotes the probability of the activation value toward the LSTI deep feature $F_{i,j}$, NT denotes the total number of adopted training instance, Alq (Eq. 4) measures the inter-frame non-local alignment reliability, $S_{i,j}$ denotes the saliency output of the j-th aligned LSTI toward the i-th superpixel, and Φ denotes the collection of all network parameters.

Meanwhile, because the spatial-temporal saliency consistency is extremely important for the high-quality video saliency detection [13], we propose to integrate it into our deep learning framework. Thus, we can formulate the temporal loss of the proposed LSTI as Eq. 8.

$$L_{tem} = \sum_{i=1:N} C(u, i)L_{non}^{t-1}(u) + C(v, i)L_{non}^{t+1}(v), \quad (8)$$

where the superscript t denotes the frame index, u, v, i are the subscripts to denote three spatial-temporally neighbored superpixels, function $C(u, i)$ returns the color similarity, which

Fig. 8. Qualitative comparisons over UCF dataset, where (a) denotes the source input video frame, and the human eye fixation based ground truths (GT) are demonstrated as the red rectangle in (a), (b) demonstrates the results obtained by our method (highlighted with red rectangle), and some state-of-the-art methods, including RADF18 [53], DLVSD18 [20], FL18 [33], FD17 [15], DSS17 [36], RFCN16 [55], DHS16 [54], MDF15 [35], GF15 [13] and SA15 [32].

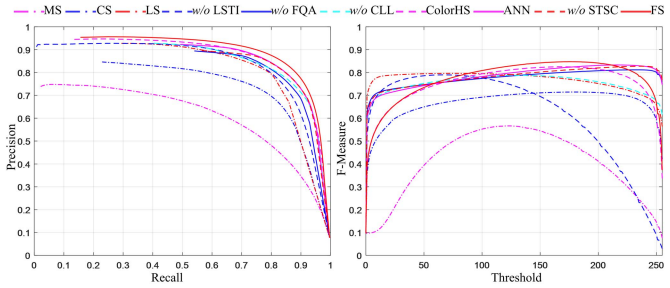

Fig. 9. Component qualitative evaluation curves over Davis [60] and SegTrackV2 [62] dataset.

can be computed by $exp(-\omega \cdot ||c_u, c_i||_2)$, where c_u denotes the RGBLab color info of the u-th superpixel, and ω is the weighting parameter. Therefore, the final spatial-temporal loss of our LSTI can be formulated as Eq. 9.

$$L_{final} = L_{non} + L_{tem}. \tag{9}$$

By minimizing L_{final}, we can simultaneously utilize the LSTI to robust the current saliency detection while ensuring the saliency consistency between consecutive video frames. And the detailed architecture of our LSTI guided deep saliency model can be found in Fig. 5.

V. Experiment and Evaluations

We have conducted massive quantitative experiments over 8 public available dataset to prove the effectiveness of our method. All video frames are down-sampled to 300 × 300 to balance the performance trade-off between the efficiency and the accuracy. The dataset adopted in our quantitative evaluation includes SegTrackV2 [62], Davis [60], DS [61], UCF [63], Visal [13], FBMS [64], UVSD [65], and MCL [31].

A. Evaluation Metrics

We adopt the F-measure and the mean absolute error (MAE) to evaluate the performance of our method, including: the F-measure and the mean absolute error (MAE). As the recall rate is inversely proportional to the precision, the tendency of the trade-off between precision and recall can truly indicate the overall video saliency detection performance. Thus, we utilize the F-measure ($\beta^2 = 0.3$) to evaluate such trade-off. Moreover, since both metrics of MAE and F-measure are based on pixel-wise errors and often ignore the structural similarities, we also adopt the structure measure S-measure [66] and enhanced-measure E-measure [67] to conduct quantitative evaluation.

B. Implementation Details

We implement our method using Matlab2016b with Caffe toolbox. We over segment all input frames into two scales, i.e., we initial assign the superpixel number as {300, 500} respectively, to handle the varying scale of the salient fore-grounds. We empirically assign the hard threshold T (Eq. 2), the weighing parameter ω of the function $C(\cdot)$ (Eq. 8), and the maximum inter-image alignment number of the long-term info m respectively as {0.3, 5, 5}. All these parameters are fixed throughout all experiments. And all evaluations are conducted on a workstation with NVIDIA GTX 1080Ti GPU, Intel Xeon W-2133 CPU (6 cores with 12 threads) and 32G RAM.

Both the optical flow computation (down-sampled to 150 × 150) and the SIFT-Flow computation (down-sampled to 100 × 100) are parallel computed by CPU, which respectively takes about 0.08s and 0.2s per frame. Also, we use CUDA to accelerate the non-local contrast based motion saliency computation, which improves the computation cost from 2s to 0.2s. Meanwhile, the deep feature computation almost

TABLE I

QUANTITATIVE COMPARISON RESULTS OF 18 STATE-OF-THE-ART METHODS OVER 8 DATASETS. BECAUSE THE CODE OF MBNM [50], SCNN [51], SCOM [52] ARE CURRENTLY NOT PUBLIC AVAILABLE, WE REFER THEIR RESULTS FROM THE PAPER [40], AND "*" DENOTES THE ABSENT DATA. THE COLUMN-WISE BESTS ARE MARKED WITH RED COLOR, THE 2ND-BESTS ARE MARKED WITH GREEN COLOR, AND THE 3RD-BESTS ARE MARKED WITH BLUE COLOR

DataSet	Metric		2018						2016-2017				2013-2015							
		LSTI Ours	FL [33]	DLVSD [20]	MBNM [50]	SCOM [52]	SCNN [51]	RADF [53]	FD [15]	DSS [36]	DHS [54]	RFCN [55]	MDF [35]	GF [13]	MC [31]	SA [32]	SU [57]	MF [58]	HS [59]	CS [60]
Davis [61]	maxF	.880	.739	.748	.861	.783	.714	.781	.758	.757	.767	.380	.720	.621	.263	.554	.261	.464	.455	.338
	maxP	.762	.599	.679	*	*	*	.723	.664	.651	.719	.387	.649	.563	.275	.582	.314	.397	.379	.320
	maxR	.922	.795	.772	*	*	*	.800	.792	.796	.782	.378	.745	.641	.259	.547	.249	.489	.484	.343
	S-M	.827	.701	.749	.887	.832	.783	.780	.706	.766	.768	.551	.679	.654	.459	.641	.599	.545	.531	.468
	E-M	.966	.865	.893	*	*	*	.872	.857	.903	.897	.776	.815	.833	.677	.826	.571	.665	.651	.431
	MAE	.031	.053	.055	.031	.048	.064	.050	.055	.049	.043	.082	.068	.099	.244	.101	.102	.178	.246	.108
DS [62]	maxF	.911	.833	.856	*	*	*	.888	.771	.810	.898	.579	.859	.478	.674	.716	.553	.598	.591	.583
	maxP	.813	.754	.772	*	*	*	.799	.643	.666	.818	.471	.713	.645	.593	.663	.503	.516	.477	.511
	maxR	.951	.860	.885	*	*	*	.918	.819	.866	.926	.622	.916	.444	.703	.733	.569	.628	.637	.608
	S-M	.857	.772	.793	*	*	*	.851	.721	.776	.854	.607	.783	.556	.605	.623	.467	.522	.556	.437
	E-M	.975	.908	.896	*	*	*	.957	.848	.874	.964	.817	.884	.767	.727	.752	.489	.687	.781	.484
	MAE	.037	.082	.057	*	*	*	.039	.063	.062	.042	.094	.069	.111	.098	.102	.122	.122	.112	.118
SegV2 [63]	maxF	.862	.831	.747	.716	.764	*	.807	.820	.679	.810	.368	.714	.739	.500	.716	.564	.275	.601	.497
	maxP	.740	.629	.616	*	*	*	.708	.660	.556	.716	.272	.514	.575	.437	.518	.450	.349	.455	.384
	maxR	.919	.915	.798	*	*	*	.841	.884	.727	.843	.412	.808	.807	.522	.809	.611	.258	.665	.546
	S-M	.827	.771	.704	.809	.815	*	.810	.750	.709	.754	.529	.638	.712	.607	.718	.485	.574	.472	.451
	E-M	.930	.873	.807	*	*	*	.901	.835	.838	.855	.614	.656	.828	.741	.859	.581	.599	.512	.497
	MAE	.028	.043	.046	.026	.030	*	.037	.037	.042	.035	.063	.050	.078	.160	.088	.125	.199	.098	.127
UCF [64]	maxF	.723	.636	.591	*	*	*	.606	.645	.567	.623	.484	.576	.571	.444	.601	.495	.499	.479	.344
	maxP	.612	.577	.407	*	*	*	.448	.523	.427	.482	.277	.413	.535	.435	.438	.400	.436	.390	.403
	maxR	.765	.656	.684	*	*	*	.678	.693	.629	.683	.625	.653	.580	.447	.676	.533	.521	.514	.329
	S-M	.554	.521	.534	*	*	*	.542	.501	.520	.531	.464	.504	.501	.441	.511	.361	.451	.438	.410
	E-M	.594	.587	.598	*	*	*	.609	.522	.574	.586	.482	.549	.571	.614	.552	.417	.511	.442	.504
	MAE	.123	.135	.152	*	*	*	.140	.138	.140	.143	.180	.156	.146	.183	.152	.167	.165	.180	.186
Visal [13]	maxF	.909	.779	.877	.883	.831	.831	.885	.784	.893	.901	.483	.822	.725	.416	.731	.687	.715	.699	.557
	maxP	.943	.821	.919	*	*	*	.918	.815	.937	.942	.514	.861	.753	.472	.768	.745	.818	.816	.622
	maxR	.836	.780	.811	*	*	*	.807	.776	.803	.847	.451	.777	.751	.415	.737	.686	.604	.557	.541
	S-M	.922	.816	.897	.898	.762	.847	.896	.822	.902	.919	.558	.851	.773	.580	.770	.684	.729	.672	.648
	E-M	.931	.790	.882	*	*	*	.907	.831	.922	.937	.812	.887	.734	.499	.715	.528	.693	.658	.606
	MAE	.027	.070	.041	.020	.022	.071	.036	.057	.033	.030	.087	.045	.099	.186	.096	.136	.156	.187	.132
FBMS [65]	maxF	.795	.594	.733	.816	.797	.762	.747	.623	.755	.736	.404	.671	.553	.222	.521	.581	.492	.454	.364
	maxP	.842	.693	.801	*	*	*	.800	.723	.817	.771	.455	.729	.613	.217	.559	.682	.540	.528	.407
	maxR	.753	.615	.658	*	*	*	.705	.594	.696	.724	.340	.663	.567	.530	.596	.548	.527	.483	.401
	S-M	.816	.655	.773	.857	.794	.794	.794	.673	.793	.789	.522	.666	.637	.503	.633	.585	.612	.543	.500
	E-M	.816	.641	.752	*	*	*	.800	.683	.810	.815	.638	.677	.629	.461	.614	.461	.584	.512	.494
	MAE	.084	.163	.105	.047	.079	.079	.095	.132	.089	.086	.154	.118	.177	.229	.185	.192	.215	.301	.239
UVSD [66]	maxF	.688	.540	.551	.550	.420	.550	.512	.554	.582	.601	.187	.467	.452	.267	.429	.472	.292	.268	.247
	maxP	.742	.625	.590	*	*	*	.533	.650	.617	.640	.199	.509	.501	.276	.482	.555	.419	.359	.298
	maxR	.654	.566	.568	*	*	*	.582	.508	.602	.601	.212	.556	.490	.404	.479	.470	.328	.229	.213
	S-M	.801	.679	.721	.698	.555	.712	.710	.710	.742	.755	.497	.669	.625	.536	.635	.591	.527	.495	.482
	E-M	.803	.690	.717	*	*	*	.721	.736	.784	.802	.644	.721	.644	.488	.630	.501	.530	.504	.512
	MAE	.037	.070	.056	.079	.206	.075	.074	.054	.047	.045	.065	.059	.131	.173	.105	.122	.193	.253	.241
MCL [31]	maxF	.680	.668	.557	.698	.422	.628	.528	.663	.567	.649	.155	.542	.391	.419	.408	.621	.292	.268	.316
	maxP	.771	.808	.621	*	*	*	.596	.773	.631	.708	.210	.599	.433	.482	.477	.736	.419	.343	.381
	maxR	.579	.568	.532	*	*	*	.639	.548	.550	.623	.181	.563	.545	.472	.453	.539	.328	.282	.341
	S-M	.783	.720	.687	.755	.569	.730	.699	.729	.710	.740	.486	.661	.603	.565	.601	.631	.527	.514	.558
	E-M	.801	.724	.658	*	*	*	.685	.754	.740	.763	.504	.715	.601	.503	.592	.524	.530	.507	.553
	MAE	.039	.049	.056	.119	.204	.054	.071	.053	.047	.032	.067	.045	.136	.167	.139	.114	.193	.242	.114

takes 0.63s per frame, and it takes an additional 0.25s for the saliency prediction. Thus, our method totally takes about 1.4s for a single video frame.

We formulate our training data from three dataset, i.e., the SegTrack2, DS, and the Davis, which totally contains about 5000 video frames with high-quality pixel-wise annotations. Since there are totally $5000(\text{frames}) \times 5(m) \times (300+500)(\text{two}-\text{scale SLIC}) = 20$ million non-local LSTI samples, where the salient samples take almost 10% of the total. Therefore, to avoid the over-fitting problem, we randomly select 30k salient samples and 30k non-salient samples to train our deep saliency model, which is approximately 0.3% of the entire dataset. We initialize the weights of new layers using gaussian distribution, and the remaining FC layers are fine-tuned on [68] proposed model. We train the network for 25k iterations, using Stochastic Gradient Descent (SGD) with a moment 0.9, weight decay 0.005, and learning rate 0.001.

C. Component Evaluation

To demonstrate the robustness of our method toward the color saliency model, we have replaced the DSS17 [36] (adopted in our previous submission) with the conventional HS13 [58], and we denote the quantitative result after adopting the HS13 as "ColorHS". By comparing the quantitative results between ColorHS and FS (our final saliency using DSS17), we can notice a slight performance degeneration after replacing the DSS17 model with the HS13 model, and we believe this is reasonable because our LSTI method is designed to further boost the detection performance over the pre-computed low-level saliency. Meanwhile, since the usage of HS13 won't

TABLE II
THE DETAILED COMPONENT QUANTITATIVE EVALUATION RESULTS

	maxF	maxP	maxR	S-M	E-M	MAE
MotionSaliency (MS)	.566	.635	.595	.625	.574	.176
ColorSaliency (CS)	.714	.754	.716	.787	.820	.046
LowlevelSaliency (LS)	.790	.853	.730	.808	.777	.043
w/o-LSTI	.796	.841	.728	.813	.869	.035
w/o-FQA	.811	.860	.747	.843	.921	.029
w/o-CLL	.826	.856	.743	.853	.927	.029
w/o-STSC	.827	.866	.759	.862	.933	.028
ANN	.833	.878	.742	.866	.926	.028
ColorHS	.829	.873	.748	.864	.897	.029
FinalSaliency	.847	.890	.773	.877	.932	.027

TABLE III
MEASURES TO ALLEVIATE LSTI ALIGNMENTS ERROR RATE OVER DIFFERENT DATA SETS. WE EXCLUDE THE DS DATASET HERE BECAUSE ITS GROUNDTRUTHES ARE COARSELY MARKED BY RECTANGLES

Dataset\Component	Frame-level Align, Eq. 4	Non-Local Align, Eq. 5	Non-Local Align with EM, Eq. 6
Davis16[61]	18%	6.6%	3.8%
DS[62]	9.9%	7.2%	3.2%
SegTrackv2[63]	12%	9.1%	3.9%

lead to a significant performance degeneration, we believe our proposed LSTI method is relatively insensitive to the adopted color saliency model.

To validate the effectiveness of our fast quality assessment (FQA), we have attempted to implement our LSTI without using the FQA component, i.e., we select multiple (with identical number to our FQA strategy) video frames with fixed time interval as the key frames, and we denote the quantitative result as "w/o-FQA". The quantitative results can be found in Table. II and Fig. 9, where the significant performance decreasing of the w/o-FQA suggests the effectiveness of our FQA strategy. As for the coarse-level localization (we abbreviate it as CLL), because its usage can effectively shrink the problem domain while alleviating the large displacement inducted incorrect SIFI-Flow alignments, the absent of CLL component decreases the overall performance (see the quantitative result "w/o-CLL"), indicating its effectiveness. Further, our spatial-temporal consistency constraint also impacts the overall performance, and we denote the performance without using this constraint as "w/o-STSC", and its corresponding quantitative results (Table. II and Fig. 9) indicate a decreased performance. We also validate the effectiveness of our long-term info usage by abandoning the LSTI alignments, which is denoted by "w/o-LSTI", and we can easily notice a significant performance decreasing of "w/o-LSTI", proving its effectiveness.

D. Compare with the State-of-the-arts

We compare the proposed algorithm with 18 state-of-the-art methods, including deep learning based methods as well as other non-deep competitors including MBNM18 [50], SCNN18 [51], SCOM18 [52], RADF18 [53], DLVSD18 [20], FL18 [33], FD17 [15], DSS17 [36], RFCN16 [55], DHS16 [54],MDF15 [35], MC15 [31], GF15 [13], SA15 [32], SU14 [56],CS13 [59], HS13 [58] and MF13 [57]. For the objective comparisons, all quantitative evaluations are conducted using the source codes provided by the authors with parameters unchanged.

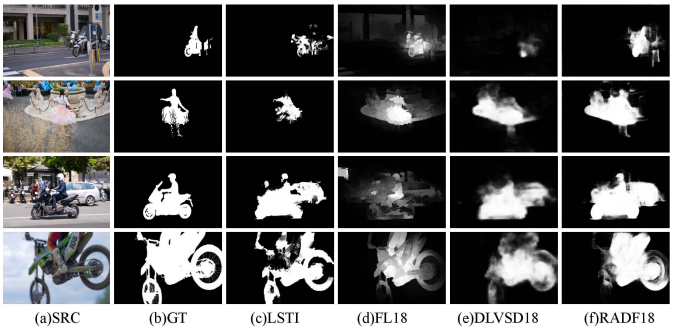

Fig. 10. The failure cases of our proposed method. Such cases are mainly caused by the weak long-term spatial-temporal coherency, which may be alleviated by formulating the LSTI within the self-paced manner.

We demonstrate the detailed F-measure, Max F-measure, MAE, as well as the corresponding precision and recall scores (maxP and maxR) in Table. I. All these quantitative results demonstrate the performance advantages comparing to the state-of-the-art methods.

For the comparison results over the Davis dataset (Fig. 7) and the UCF dataset (Fig. 8), almost all the current state-of-the-art methods easily produce massive false-alarm detections for those sequences with dynamic backgrounds while coupling irregular movement patterns, and the hollow effects can also be frequently observed. However, benefiting from our newly introduced LSTI, such failure detections can correctly be solved by our method. Meanwhile, as for the SegTrackV2 and Visal datasets, our method can significantly outperform in recall rate, because these datasets are dominated by motions, and those short-term motion abruptiones can be well handled by our non-local LSTI strategy.

Specially, as for the quantitative comparisons over the DS dataset, our method can only slightly outperform the other methods, which is mainly because the DS dataset is dominated by the color info, and the solely spatial info based image saliency detection methods (e.g., DHS16) can also produce competitive detections. Thus, we believe that our performance over DS dataset can be potentially boosted by adopting more accurate color saliency network.

E. Limitations

Because our LSTI revealing is heavily dependent on those beyond scope long-term info, our method can only be operated within the off-line manner. Another limitation is that our method tends to be time-consuming, i.e., almost 1.4s for a single 300×300 video frame. Consequently, our method temporally not supports the end-to-end real-time video saliency detection. Also, the effectiveness if our LSTI scheme toward the performance improvements may be limited in those videos with weak spatial-temporal coherency, e.g., with rapid and

TABLE IV
THE DETAILED AVERAGE TIME CONSUMPTION TOWARD DIFFERENT PARTS OF THE PROPOSED METHOD

MainSteps	Time(seconds)
Optical Flow Computation	0.090
SLIC Superpixel	0.050
Motion Saliency	0.040
Color Saliency	0.109
Fast Quality Assessment (Sec. III-B)	0.040
Coarse-level Localization (Sec. III-D)	0.020
Fine-level SIFT-Flow (Sec. III-C)	0.500
Deep Feature Computation (Sec. IV-A)	0.410
Saliency Prediction (Sec. IV-C)	0.100
Total	1.359

drastic scale variations, intermittent occlusion and revealment, which easily increases the error rate of the aligned LSTI. And the failure cases can be found in Fig. 10.

VI. CONCLUSION

This paper has proposed a novel supervised deep convolutional network for video saliency detection, which utilizes the long-term spatial-temporal information (LSTI) to facilitate saliency detection. Due to the absence of long-term information, both the intermittent movements induced hollow effects and the external disturbance caused false-alarm detections can frequently occur in the conventional methods using only short-term spatial-temporal information. Our new method intends to reveal the LSTI from the high-quality low-level saliency estimations, which could be characterized using the newly designed fast quality assessment (FQA) scheme, by performing non-local inter-frame alignments guided by SIFT-Flow. Next, we utilize a novel deep saliency framework to take full advantage of the newly available LSTI to simultaneously learn the discriminative information towards the salient foregrounds while maintaining strong spatial-temporal saliency consistency in order to achieve high-performance video saliency detection. In our near-future works, we will focus on two aspects. First, we will endeavor to adapt our method into an end-to-end architecture to enable real-time video saliency detection. Second, by using GANs to generate labeled training datasets, we are able to alleviate the training data shortage dilemma towards the end-to-end model training in video saliency.

REFERENCES

[1] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[2] G. Ma, C. Chen, S. Li, C. Peng, H. Qin, and A. Hao, "Salient object detection via multiple instance joint re-learning," *IEEE Trans. Multimedia*, to be published.

[3] W. Wang, Q. Lai, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[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] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.

[7] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4225–4232.

[8] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

[9] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognit.*, vol. 48, no. 9, pp. 2885–2905, 2015.

[10] 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.

[11] C. Peng, C. Chen, Z. Kang, J. Li, and Q. Cheng, "RES-PCA: A scalable approach to recovering low-rank matrices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7317–7325.

[12] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *Proc. Int. Conf. Inf. Sci. Cloud Comput. Companion*, Dec. 2013, pp. 728–733.

[13] 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.

[14] L. Zhi, 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.

[15] 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.

[16] 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.

[17] C. Chen, G. Wang, and C. Peng, "Structure-aware adaptive diffusion for video saliency detection," *IEEE Access*, vol. 7, pp. 79770–79782, 2019.

[18] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 186–202.

[19] L. Jiang, M. Xu, and Z. Wang, "Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM," in *Proc. Eur. Conf. Comput. Vis.*, Dec. 2018, pp. 1–16.

[20] 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.

[21] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3243–3252.

[22] R. Jiang and D. Crookes, "Deep salience: Visual salience modeling via deep belief propagation," in *Proc. 28th AAAI Conf. Artif. Intell.*, Jul. 2014, pp. 2773–2779.

[23] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4894–4903.

[24] 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.

[25] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

[26] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[27] J. Shen, J. Peng, and L. Shao, "Submodular trajectories for better motion segmentation in videos," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2688–2700, Jun. 2018.

[28] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 985–998, Apr. 2019.

[29] Z. Feng, S. Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3358–3365.

[30] D. Zhang, J. Omar, and S. Mubarak, "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.

[31] 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.
[32] 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.
[33] 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.
[34] X. Zhou, Z. Liu, C. Gong, and W. Liu, "Improving video saliency detection via localized estimation and spatiotemporal refinement," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2993–3007, Nov. 2018.
[35] 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.
[36] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.
[37] T.-N. Le and A. Sugimoto, "Video salient object detection using spatiotemporal deep features," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5002–5015, Oct. 2018.
[38] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 715–731.
[39] S. D. Jain, B. Xiong, and K. Grauman, "FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2126.
[40] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8544–8564.
[41] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
[42] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.
[43] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.
[44] C. Liu *et al.*, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Dept. Sci. Technol., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.
[45] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.
[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[47] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
[48] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.
[49] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[50] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. J. Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 207–223.
[51] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, and X. Li, "Weakly supervised salient object detection with spatiotemporal cascade neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1973–1984, Jul. 2019.
[52] Y. Chenm, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, "SCOM: Spatiotemporal constrained optimization for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3345–3357, Jul. 2018.
[53] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6943–6950.
[54] 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.
[55] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 825–841.
[56] Y. Fang, Z. Wang, and W. Lin, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2013, pp. 1–6.
[57] 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.
[58] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, Apr. 2016.
[59] 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.
[60] 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.
[61] K. Fukuchi, 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.
[62] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2192–2199.
[63] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *Proc. Eur. Conf. Comput. Vis.* Florence, Italy: Springer, 2012, pp. 842–856.
[64] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.
[65] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.
[66] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4548–4557.
[67] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.
[68] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3183–3192.

Chenglizhao Chen received the Ph.D. degree in computer science from Beihang University in 2017. He is currently an Assistant Professor with Qingdao University. His research interests include computer vision, machine learning, and pattern recognition.

Guotao Wang received the B.S. degree in computer science from Qingdao Agricultural University in 2017. He is currently pursuing the master's degree with Qingdao University. His research interests include computer vision and deep learning.

Chong Peng received the Ph.D. degree in computer science from Southern Illinois University Carbondale, Carbondale, IL, USA, in 2017. He is currently an Assistant Professor with Qingdao University. His research interests include machine learning, pattern recognition, and data mining.

Xiaowei Zhang received the Ph.D. degree in computer science from Beihang University in 2018. He is currently an Assistant Professor with Qingdao University. His research interests include computer vision, machine learning, and pattern recognition.

Hong Qin (SM'08) received the B.S. and M.S. degrees from Peking University and the Ph.D. degree from the University of Toronto, all in computer science. He is currently a Professor of computer science with the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing.