Pattern Recognition 100 (2020) 107158

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Joint temporal context exploitation and active learning for video segmentation

Yan Tian[a,b], Guohua Cheng[c], Judith Gelernter[d], Shihao Yu[a], Chao Song[a], Bailin Yang[a,*]

[a] School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, PR China
[b] Shining3D Research, Shining3D Tech Co., Ltd., Hangzhou 310018, PR China
[c] Institute of Science and Technology for Brain-Inspired Intelligence, Ministry of Education-Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai 200433, PR China
[d] Information Science Department, Rutgers University, New Brunswick 08901, USA

ARTICLE INFO

Article history:
Received 29 May 2019
Revised 29 October 2019
Accepted 11 December 2019
Available online 14 December 2019

Keywords:
Video segmentation
Deep learning
Computer vision

ABSTRACT

The segmentation of video, or separating out objects in the foreground, is an important application of pattern recognition and computer vision. Segmentation errors in pattern recognition approaches mainly come from difficulties in selecting maximally informative frames for learning. In this paper, we develop an approach to video segmentation that relies on temporal features by modeling the uncertainty of the distribution of different feature mask forms. We use those uncertainty values for unsupervised active learning. We evaluate our approach on the DAVIS16 annotated video data set and Shining3D dental video data set, and the results show our approach to be more accurate than other video segmentation approaches.

1. Introduction

Video segmentation usually aims to separate, or segment, some objects in the foreground from the video background. This topic has been an active area of research in pattern recognition over the past decade, with a wide range of applications. Examples of applications include video editing [1], video summarization [2], scene understanding [3], and autonomous driving [4].

Recently, methods such as one-shot learning [5] and meta-learning [6] have been applied to video segmentation. However, progress in image segmentation has been limited by temporal contexts such as feature alignment, mask propagation, and motion exploitation [7]. This is because these temporal inferences cannot be compared with spatial inferences, which leads to ambiguity in determining where to segment an image.

We illustrate the video segmentation problem in Fig. 1. Here, the leftmost image is the input, the second and third images are the corresponding spatial (DeepLabV3+ [8]) and temporal (MaskTrack [9]) inference results, and the fourth image is the ground truth. Careful inspection of the second and third images shows that segmentation from spatial and temporal methods is very good, but not perfect. Spatial and temporal inference have their own advantages in different situations.

Objects are segmented even less well than those in Fig. 1 when the video has a cluttered background or dim lighting. Such harder samples use boundary knowledge to guide the refinement. However, harder samples can only be distinguished with the aid of human resources or a time-consuming training process in machine learning [10]. Video segmentation requires an unsupervised active learning approach to identify the representative samples. To this end, we determine the uncertainty of samples by measuring the confidence value produced by spatial and temporal inference. We use this uncertainty to select the most informative samples for active learning.

In this paper, we develop a probability-based approach that uses a temporal inference such as optical flow to model the distribution in mask propagation for sample selection.

Our approach is demonstrated in Fig. 2. Image segmentation and motion inference are learned simultaneously in multitask learning. This approach is used to select an image mask, which is transferred to the next frame in a probability framework. Finally, samples with maximum uncertainty are selected. This process improves the accuracy of active learning.

The contributions of this paper are the following:

- We introduce multitask learning using a recurrent network to improve the accuracy of motion inference in multiscale analysis.
- We present a novel way to use temporal context to infer the mask in the next frame via a probability framework, and

[*] Corresponding author.
E-mail address: ybl@zjgsu.edu.cn (B. Yang).

https://doi.org/10.1016/j.patcog.2019.107158
0031-3203/

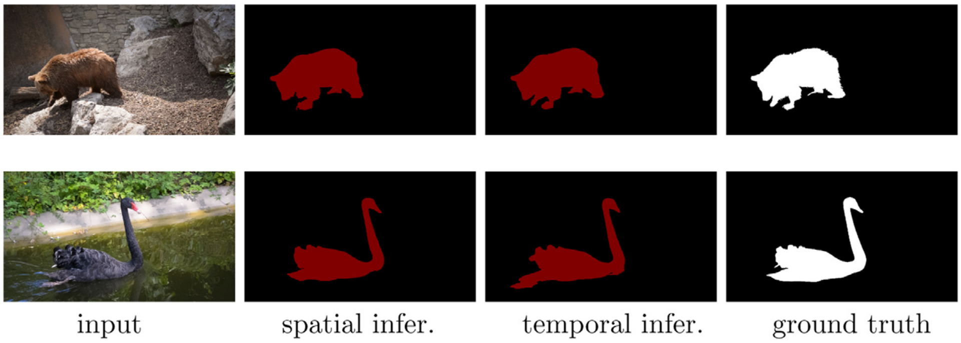


Fig. 1. Illustrations of spatial and temporal inference in the DAVIS16 data set. The first column is the input, the second and third columns are spatial and temporal inference results, and the fourth column is the corresponding ground truth.

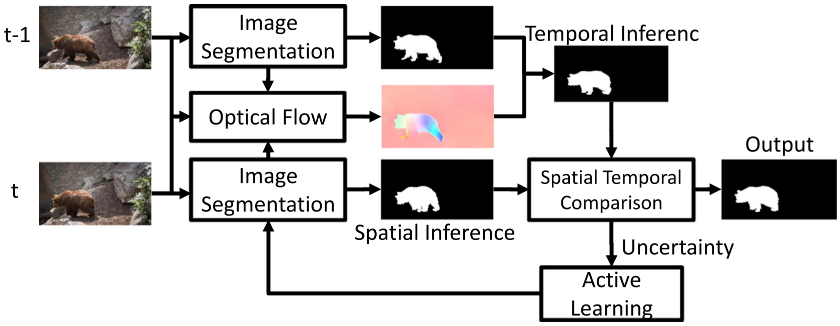


Fig. 2. Example of how our approach works. Image segmentation and motion inference are estimated jointly, and the result is refined by means of the spatial-temporal comparison. Each segment is produced with an uncertainty value that informs segment selection over time via active learning..

we compare this result to the spatial result according to the confidence interval.

- We use uncertainty in the probability mask transfer to improve sample selection for unsupervised active learning.
- We have collected and annotated an oral video sequence from the Shining3D dental data set so that the approach proposed in this paper can be compared to other state-of-the-art approaches on video segmentation. We have published this data set so that anyone working on this topic may freely obtain this resource.

We show in the Results section that, in experiments with the DAVIS16 data set and the Shining3D dental data set, our approach yields results that are competitive with those of state-of-the-art video segmentation approaches.

2. Related work

In this section, we briefly review the literature on the use of temporal features for video segmentation and active learning. We present advantages and drawbacks of each approach.

2.1. Video segmentation

Generally, video segmentation can be divided into two different categories, namely unsupervised and semi-supervised methods.

Unsupervised video segmentation does not provide any manual annotation. To explore motion information from a long time span, pyramid dilated bidirectional ConvLSTM (PDB-ConvLSTM) [11] and a two-stream neural network [12] model spatial-temporal features on a convolutional LSTM and convolutional GRU structure, respectively. Wang et al. later introduce temporal and spatial attention into this task [13], using visual stimuli to guide unsupervised video object segmentation. Considering that primary objects tend to be highly correlated at the macro level, a co-attention Siamese network (COSNet) [14] learns to capture rich correlations between frames in a video sequence.

Semi-supervised video segmentation provides information about objects of interest in the first frame of a video. Spatiotemporal Markov random field (MRF) [15] models spatial dependencies among pixels by a convolutional neural network and establishes temporal dependencies by optical flow. To adapt to large changes in object appearance in a video, online adaptive video object segmentation (OnAVOS) [16] updates the network online using training examples selected based on the confidence of the network and the spatial configuration. Sometimes temporal smoothness will be suddenly broken; therefore, semantic one-shot video object segmentation (OSVOS-S) [17] has been proposed to process each frame independently, i.e., disregarding the temporal information. The PReMVOS algorithm (proposal-generation, refinement and merging for video object segmentation) [18] considers multiple factors and combines an objectness score, optical flow warping, and a Re-ID feature embedding to warp the segmented mask to the next frame. Nevertheless, these approaches are limited by suboptimal accuracy due to mismatching and drifting problems; ranking attention network (RANet) [19] automatically ranks and selects these maps for fine-grained video object segmentation performance.

These approaches return inference results by using temporal consistency. However, the uncertainty of the spatial and temporal inference cannot be compared, which leads to ambiguity when the segment from each of these two approaches does not match.

2.2. Active learning

Active learning is a form of machine learning in which the human user selects representative samples that inform the learning process. In deep learning, dropout [20] has been used to reduce overfitting and improve model generalizability [21]. The probability output can be obtained directly by a Bayesian convolutional neural network [10] to help represent model uncertainty. To increase sample diversity in the instances, clustering and fuzzy-set selection have been used to select representative and informative samples [22], as well as two-sample discrepancy [23]. These approaches require a time-consuming training process. In short, identification of the best samples for active learning or deep learning, especially when the data are high-dimensional images, is a complex process.

Recently, an unsupervised method to identify informative samples by means of template matching in neighbor frames has been proposed [24]. Some researchers have modified this approach by using motion and saliency to identify negative samples. Nevertheless, this approach uses only the intersection or union of samples found by spatial and temporal inference.

2.3. Saliency detection

Saliency detection aims to automatically discover and locate regions that are visually interesting. The focus of salient object detection is on designing various computational models to measure image saliency, which is useful for segmentation. The contour-to-saliency network (C2S-Net) [25] learns contour-to-saliency and saliency-to-contour iteratively. Li et al. [26] set up a multi-task learning scheme for exploring intrinsic correlations between saliency detection and semantic image segmentation. The multi-scale bidirectional fully convolutional network (MSBFCN) [27] generates feature maps of different scales directly from the last convolutional layer of a pretrained underlying model by using a pyramid pooling strategy; then, a bidirectional structure is employed to capture and encode multi-context information. The multiscale cascade network (MSC-Net) [28] enables the learning process in the finer cascade stages to encode more global contextual information, while knowledge is obtained progressively incorporating the saliency information. In this way, the approach leads to better detection accuracy. Li et al. [29] transfer annotations from an existing example onto an input image. Wang et al. [30] propose a deep learning model composed of two modules that are designed for capturing the spatial and temporal saliency information simultaneously.

3. Our approach

To exploit temporal context most effectively for active learning, we propose the approach shown in Fig. 2. Image segmentation and motion inference are estimated jointly, and the result is refined by means of the spatial–temporal comparison. Each segment is produced with an uncertainty value that informs segment selection over time via active learning.

3.1. Joint learning of image segmentation and optical flow inference

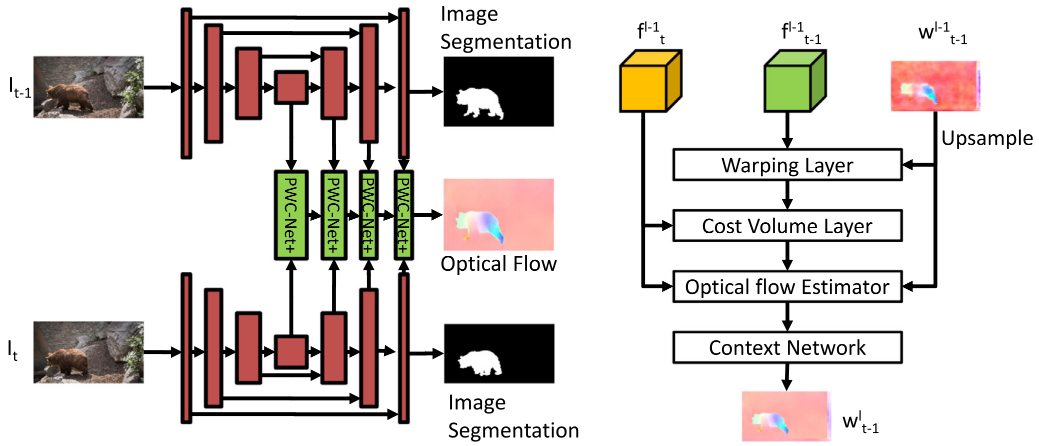
In some recent video segmentation approaches, binary masks or feature maps are propagated over time according to the optical flow obtained by using the CNN method [18]. However, the image is segmented independently of the optical flow inference, which tends to trap the inference result in the local optimum.

Image segmentation and inferences from optical flow are related tasks. On the one hand, differences in binary masks among neighbor frames give clues for motion estimation; on the other hand, it is the motion that causes the segmentation result to differ in successive frames. Therefore, we design a mechanism to jointly learn the image segmentation and optical flow inference, as illustrated in Fig. 3(a). The segmented mask is obtained by using the hierarchy residual learning subnetwork [31,32]. To effectively and efficiently segment a video sequence, our approach is based on the accelerated RefineNet [33] and PWC-Net+ because these two approaches both use hierarchical analysis and are combined easily.

Our image segmentation and motion estimation share the encoder network for model compression. The decoder layers during segmentation are fused with optical flow estimators. We use the decoder part for fusion because this part contains information that is semantic and also high in resolution detail.

Here is a formal, mathematical description of our approach, as shown in Fig. 3(b). Given two images $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$ in $t-1$th and tth frames of the sequence, we generate feature representations $\mathbf{f}_{t-1}^{l}$ and $\mathbf{f}_{t}^{l}$ in the lth layer. We want to progressively estimate optical flow $\mathbf{w}_{t-1}^{l}$ by using these multiscale feature maps. First, we employ the warp layer to estimate large motion from one frame to the next. At the lth layer, we warp features of the image $\mathbf{I}_{t-1}$ toward the image $\mathbf{I}_t$ by using

$$\mathbf{f}_{w}^{l}(x) = \mathbf{f}_{t-1}^{l}\left(x + up\left(\mathbf{w}_{t-1}^{l-1}\right)\right), \qquad (1)$$



(a) Multi-task learning (b) PWC-Net+

Fig. 3. Illustrations of joint learning of image segmentation and optical flow inference. (a) The multitask learning framework. + The segmented mask is obtained by using the hierarchy residual learning subnetwork, and the optical flow is gained via the PWC-Net+ subnetwork. (b) Detail of PWC-Net+ block, which includes warping layer, cost volume layer, optical flow estimator, and context network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

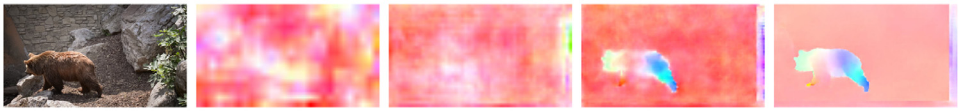


Fig. 4. Illustrations of optical flow inference. The left image is the original image, and the remaining images are optical flow images obtained from different layers in the motion inference block.

where x is the pixel index, and up is the flow upsampling operation such as a bilinear interpolation of gradient propagation.

Because cost volume is a discriminative representation of optical flow, our network constructs the cost volume $\mathbf{m}^l$ by using warped features with corresponding features in the next frame to store matching costs.

$$\mathbf{m}^l(x_1, x_2) = \frac{1}{N}\left(\mathbf{f}_w^l(x_1)\right)^{\mathrm{T}}\mathbf{f}_t^l(x_2), \tag{2}$$

where x_1 and x_2 represent the pixel index in the warped features $\mathbf{f}_w^l$ and features $\mathbf{f}_t^l$, respectively, T is the transpose operator, and N is the length of the column vector $\mathbf{f}_w^l(x_1)$.

Next, DenseNet [34] is employed to estimate the flow. The input includes the cost volume $\mathbf{m}^l$, image features $\mathbf{f}_t^l$ in frame t and layer l, and upsampled optical flow $up(\mathbf{w}_{t-1}^{l-1})$. The output is the original estimated flow $\mathbf{w}_{t-1}^l$ at the lth layer.

Finally, we use a context network. This approach uses four dilated convolution layers to exploit the optical flow. The spatial kernel for each convolutional layer is 3 × 3, while the dilation constants are 1, 2, 4, and 8.

Examples of the inferred optical flow in the decoder can be seen in Fig. 4. From this image, it can be found that the motion information is progressively made more accurate in the inference stage by combining the coarse flow in the last layer and the image feature in the current layer.

Our approach has several advantages. For nontranslational motion, warping can compensate for some geometric distortions and put objects at the right scale. Moreover, the warping and cost volume layers have no learnable parameters that serve to reduce the model size and make the remaining layers computationally light. All the blocks in the network are differential, and the image segmentation and optical flow inference can be jointly learned in an end-to-end manner so that results do not become trapped in the local minimum.

3.2. Temporal context exploitation

Image segmentation does not return results that are consistently good, especially when the object is spatially occluded or truncated. Some experts have tried to improve results using a mask propagation method, such as maskTrack. However, the task of finding corresponding masks based on temporal inference is sensitive to changes in the reflecting properties of surfaces, as well as in the presence of multiple objects moving in the scene.

Our own research centers on whether we can use both spatial and temporal inference to model uncertainty in selecting masks. Spatial inference methods give each pixel in the mask map a confidence score. If temporal propagation could be modeled similarly, then the spatial and temporal inference could be combined in a probability framework. We present this mathematically as follows:

Suppose $\mathbf{d}_s$ and $\mathbf{d}_t$ are spatial and temporal derivatives of the image $\mathbf{I}$, and $\mathbf{w}$ is the optical flow. Then, Bayes' rule gives

$$P(\mathbf{w}|\mathbf{d}_s, \mathbf{d}_t) = \frac{P(\mathbf{d}_t|\mathbf{d}_s, \mathbf{w})P(\mathbf{w})}{P(\mathbf{d}_t)}, \tag{3}$$

where $P(\mathbf{w})$ is the prior distribution that is modeled by a zero-mean Gaussian distribution with covariance Λ_p, and $P(\mathbf{w}|\mathbf{d}_s, \mathbf{d}_t)$ can be modeled by the revised optical flow equation

$$\mathbf{d}_s(\mathbf{w} - \mathbf{n}_1) + \mathbf{d}_t = \mathbf{n}_2, \qquad \mathbf{n}_i = N(0, \Lambda_i). \tag{4}$$

We use noise term $\mathbf{n}_1$ to represent errors resulting from a failure of the linearity assumptions and $\mathbf{n}_2$ to describe the errors in temporal derivative measurements. Suppose the intermediate vectors are

$$\mathbf{M} = \mathbf{d}_s\mathbf{d}_s^{\mathrm{T}}, \tag{5}$$

$$\mathbf{b} = \mathbf{d}_s\mathbf{d}_t^{\mathrm{T}}. \tag{6}$$

If Λ_1 is a diagonal matrix with diagonal entry σ_1 and the scalar variance of $\mathbf{n}_2$ is $\sigma_2 \equiv \Lambda_2$, then the posterior distribution $P(\mathbf{w}|\mathbf{d}_s, \mathbf{d}_t)$ can be described by the mean and covariance

$$\mu_w = -\Lambda_w \frac{\mathbf{b}}{\sigma_1||\mathbf{d}_s||^2 + \sigma_2}, \tag{7}$$

$$\Lambda_w = \left[\frac{\mathbf{M}}{\sigma_1||\mathbf{d}_s||^2 + \sigma_2} + \Lambda_p\right]^{-1}. \tag{8}$$

Finally, suppose the spatial inference in frame t obtains mask map $\mathbf{m}_t$ and confidence map $\mathbf{c}_t$; then, the spatial inference $\mathbf{m}_t$ in frame t and temporal inference $P(\mathbf{w}|\mathbf{d}_s, \mathbf{d}_t)$ can be combined to obtain the final segmentation $\tilde{\mathbf{m}}_t$ according to their confidence scores

$$\tilde{\mathbf{m}}_t^{x1} = \begin{cases} \mathbf{m}_t^{x1} & \text{if } \mathbf{c}_t \geq \mathbf{c}_{t-1}P(\mathbf{w}_{x1,x2}|\mathbf{d}_s, \mathbf{d}_t), \\ \mathbf{m}_{t-1}^{x2} & \text{otherwise,} \end{cases} \tag{9}$$

where x_1 and x_2 are the pixel index in posterior distribution $P(\mathbf{w}_{x1,x2}|\mathbf{d}_s, \mathbf{d}_t)$. If the spatial inference meets the temporal inference, the two types of inference obtain the same results; otherwise, the final prediction result is determined by the confidence (probability) of each type of prediction.

This formulation assumes only the loose constraint that changes in the image intensity are due to translation of the local image intensity. Then, the spatial inference in different frames can be compared according to confidence scores from the inference output.

3.3. Temporal-based, unsupervised active learning

If we have a method to measure the spatial-temporal comparison and find those samples that are prone to error, the effectiveness of the active learning could be improved.

Active learning and unsupervised hard sample mining are designed to select the most informative, diverse samples to be labeled. Typically, samples with maximum uncertainly are selected according to entropy values. However, temporal information is neglected in this uncertainty estimation.

Our framework provides spatial and temporal inference, as described in the preceding subsection. Therefore, we can use the compatibility between spatial and temporal samples to measure the inference uncertainty and refine the multitask network.

To do this, we employ the intersection-over-union (IoU) to determine positive or negative samples. If the IoU between the fusion mask $\tilde{\mathbf{m}}_t$ and spatial inference $\mathbf{m}_t$ is above a threshold T_1, we regard the segmentation in the current frame as a positive sample. We compare the image segment and the fusion mask rather than temporal inference because warped masks inevitably have some invalid region.

It is logical that the hard negatives should be selected for retraining. In practice, we employ block matching to the previous k frames of an enlarged bounding box by $20 \times k$ pixels in different directions. If the overlap between the matched bounding box P_{t-k} in the $(t-k)$-th frame and the outside bounding box of the fusion mask $f(\tilde{\mathbf{m}}_t)$ in frame t is above a threshold T_2, we denote these masks as examples of hard negatives, which can be represented as follows:

$$HN = \tilde{\mathbf{m}}_t\{if\ \ (\mathbf{m}_t \cap \tilde{\mathbf{m}}_t < \mathrm{T}_1)\ \text{and} \\ \Big(min_{k\in[1,3]} f(\tilde{\mathbf{m}}_t) \cap P_{t-k} \geq \mathrm{T}_2\Big)\Big\}. \tag{10}$$

By using this definition of hard negative samples, the representative and informative samples can be selected automatically, and improvements can be achieved by fine-tuning the network.

4. Experimental results

In this section, we compare the performance of our proposed approach to other approaches.

4.1. Hardware and software environment

We use a workstation with an Intel i7-4790 3.6 GHz CPU, 32GB memory, and NVIDIA GTX Titan X graphics. Our algorithm to verify performance and computational efficiency is based on Tensorflow [35].

4.2. Data sets

We verify our proposed approach on the DAVIS16 data set [36] and the Shining3D dental data set [37].

4.2.1. DAVIS16 data set

The DAVIS16 (Densely Annotative Video, at www.davischallenge.org) data set is composed of 50 high-resolution videos, 30 for training and 20 for validation. Those videos divide into 3455 image frames with corresponding pixel-wise annotations. The data set is used widely in video segmentation research because it contains challenges such as motion blur, occlusions, and changes in the appearance of the main object. Only the primary moving objects are annotated in the image frames of the ground truth.

4.2.2. Shining3D dental data set

Shining3D is a set of 47 nonannotated videos of human mouths generated by a 3D dental scanning device that is often used for research. Each video came from a hospital patient who was selected randomly. We made some changes to ensure that privacy would be maintained. We randomly selected and annotated 7800 images from these videos consisting of a training set of 5800 images from 40 people and a validation set of 2000 images from the remaining 7 people. The image size is fixed at 640 pixels in width and 480 pixels in height. Four researchers from our university annotated the training and validation images; another 4 researchers re-annotated the same images to ensure correctness. When the labels disagreed, we employed another person for evaluation. Each annotator used software called "LabelM" to mark the boundary and classify each region as tooth, gum, jaw, lip, cheek, or other soft tissues. For application purposes, we set teeth and gums as classes of interest and all other soft tissues as classes that were less relevant.

4.3. Evaluation criteria

We use evaluation criteria as well as data sets that others have used in published research to compare our work to state-of-the-art approaches.

The Jaccard index J is defined as the IoU between the ground truth mask and the predicted mask to measure region-based segmentation similarity. Specifically, given a predicted mask P and corresponding ground truth mask G, J is defined as $J = \frac{P\cap G}{P\cup G}$.

4.4. Implementation details

In our implementation, image segmentation is initialized by RefineNet [33], and optical flow inference is initialized with PWC-Net+ [32]. The whole network is trained on a stochastic gradient descent (SGD) method with a momentum of 0.9 and a weight decay of 0.0005. To avoid shocks in the performance curve, the learning rate is set to 0.01 for the first 60,000 iterations and 0.001 for the later 20,000 iterations, which is adjusted according to the performance of the evaluation set.

In the active learning experiments, the selection of k depends on the object or camera movement intensity. If the video contains large extent motion, the similarity between neighbor frames is small, and the matching process works on a small number of frames, that is, k is small. Otherwise, k can be increased to match the smoothness of movement. Therefore, we choose $k = 3$ in the DAVIS16 data set (small motion) and $k = 2$ in the Shining3D data set (large motion). If T_1 decreases, the number of hard samples will also decrease, which leads to unstable results owing to the lack of training data. However, if T_1 increases too much, not only the hard samples but also samples with limited value are added into the training phase, which decreases the training efficiency and brings no effectiveness benefit. In the DAVIS16 data set and Shining3D data set, we set T_1 to be 0.7, which is selected according to grid searching. T_2 measures the motion extent and has a similar function as that of k. As we already choose different k values to represent motion variation, we choose $T_2 = 0.7$ in the DAVIS16 data set and Shining3D data set.

4.5. Ablation study

We perform extensive ablation studies to observe the effects of several important components of our approach. Temporal context experiments are performed on the DAVIS16 data set only. Active learning experiments are performed on the Shining3D dental data set because it conveniently includes 130,000 images that are not annotated and thus can be labeled according to the output of the active learning.

Question 1: How many feature scale sizes should be used in the joint learning of image segmentation and optical flow inference in multiscale analysis? To answer this question, schemes with various numbers of scales were tested, and the performance comparison is in Fig. 5. We know that if more scales are used in the scheme, the multiscale analysis will improve. However, larger size networks tend to lead to overfitting. In our experiment, the use of 4 scale sizes shows the best performance in the statistical analysis. Hence, in the next experiments, we use a scheme with 4 sizes of scales to compare the effectiveness.

Question 2: Does the use of temporal mask propagation using a probability-based method improve video segmentation performance? We compare our approach to other temporal propagation approaches using the DAVIS16 validation set. The resulting performance curve (mean J) can be seen in Fig. 6. Regardless of the

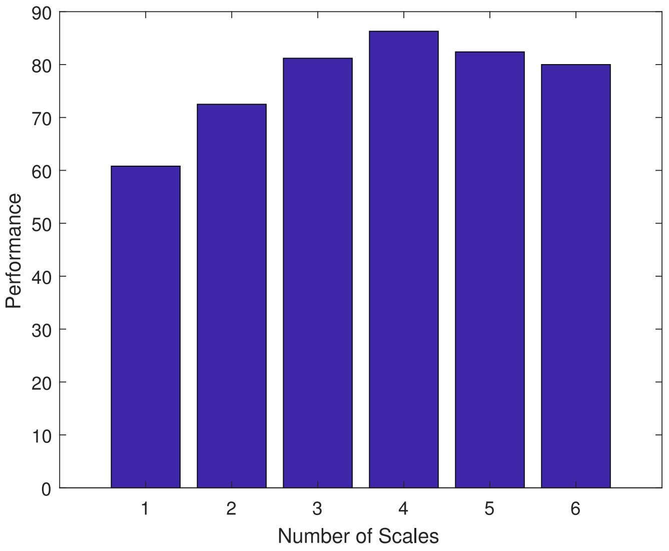


Fig. 5. Evaluation of multiscale analysis on the DAVIS16 validation set.

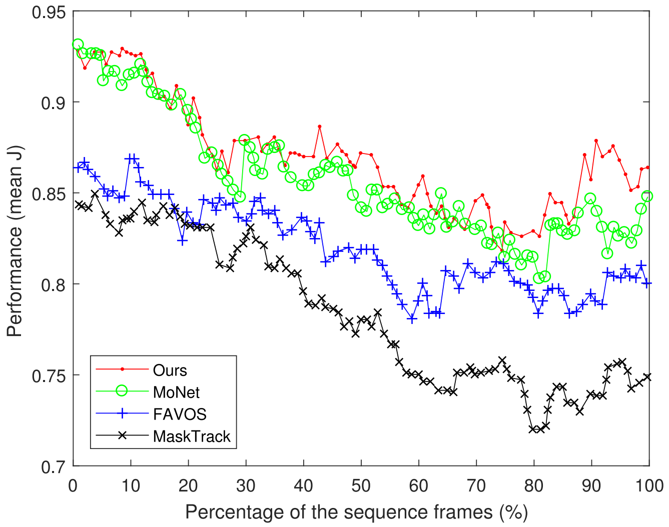


Fig. 6. Performance (mean *J*) variations of different methods over time on the DAVIS16 validation set.

method used, the figure shows that performance decreases until approximately 75% of the sequence has run and then increases again because the appearance changes compared to that in the first frame and the target object becomes harder to segment in later frames. Our probability temporal context exploitation achieves better performance than that of MoNet through the whole sequence and obtains an improvement of 1.6 in the Jaccard index by the end of the sequence, which demonstrates that our approach can effectively incorporate spatial and temporal context for video segmentation.

Question 3: Is active learning an effective approach for video segmentation? If so, does our active learning framework select more informative samples than do random-based or entropy-based approaches? The experimental results are given in Fig. 7. This figure shows that all the active learning improves the performance by a margin in terms of the Jaccard index, which demonstrates the effectiveness of active learning for video segmentation. In addition, DOA and our approach are temporal context exploitation-based active learning approaches, and they show more improvement than do the random-based and entropy-based approaches.

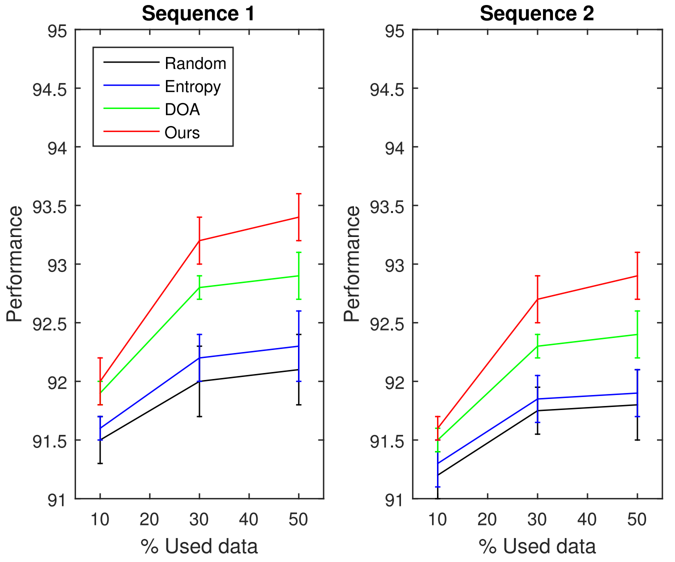


Fig. 7. Performance (mean *J*) variations of active learning methods in sequences 1 and 2 on the Shining3D dental set.

Table 1
Experiment results on the DAVIS16 validation set and Shining3D dental data set.

Approach	DAVIS16	Shining3D
ATEN [38]	71.3	80.8
MaskTrack [9]	74.8	82.0
S2S [39]	79.1	86.7
DSE [40]	81.5	88.9
MoNet [41]	84.7	91.5
PReMVOS [18]	85.5	92.0
Ours	**86.4**	**93.5**

4.6. Evaluation on the DAVIS16 validation set

We compare our approach to other temporal context exploitation approaches in video segmentation, with the experimental results shown in Table 1. Feature alignment approaches such as adaptive temporal encoding network (ATEN), sequence-to-sequence network (S2S), and MoNet attain accuracy comparable to that of mask propagation approaches, such as MaskTrack, deep Siamese encoder-decoder, and PReMVOS.

In feature alignment, recently, recurrent neural networks have been employed to effectively use temporal context. Take ATEN for example, where a gated recurrent unit (GRU) [42] is used to infer temporal encoding for key frames. To solve the problem that temporal dependency relies on optical flow and cannot be trained in an end-to-end way, a sequence-to-sequence (S2S) network uses long-term spatial temporal context with the help of long short-term memory (LSTM) [43]. The results show improvement by a Jaccard index value of 7.8 when compared with that of the ATEN.

MoNet also learns the segmentation and optical flow jointly, but the inferred optical flow is used for feature alignment. It obtains a Jaccard index value of 84.7.

MaskTrack directly learns the mapping between masks in neighbor frames, but its performance is limited by large motions of the object and intra-variance of the object's appearance. The deep Siamese encoder-decoder (DSE) combines frame information with residual learning to infer the propagated mask, and experiments show that compared to MaskTrack, it can obtain an improvement in the Jaccard index of 6.7 on the DAVIS16 data set. Recently, PReMVOS has employed a state-of-the-art image segmentation approach (DeepLabv3+) and used a re-identification approach for tracking, obtaining a Jaccard index value of 85.5 on the DAVIS16

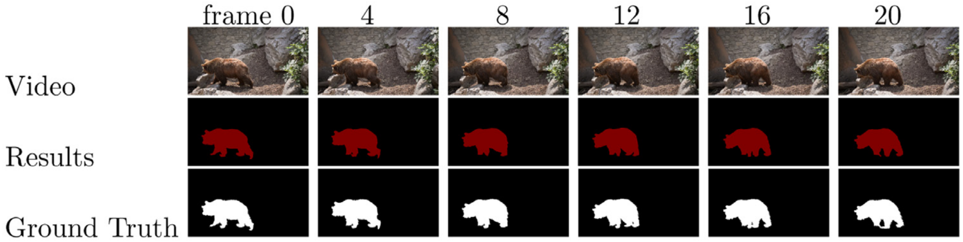


Fig. 8. Experimental results on the DAVIS16 data set. The upper row is the input video (No. 0, 4, 8, 12, 16 and 20 frames in the bear sequence), the middle row is our segmentation result, and the lower row is the corresponding ground truth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

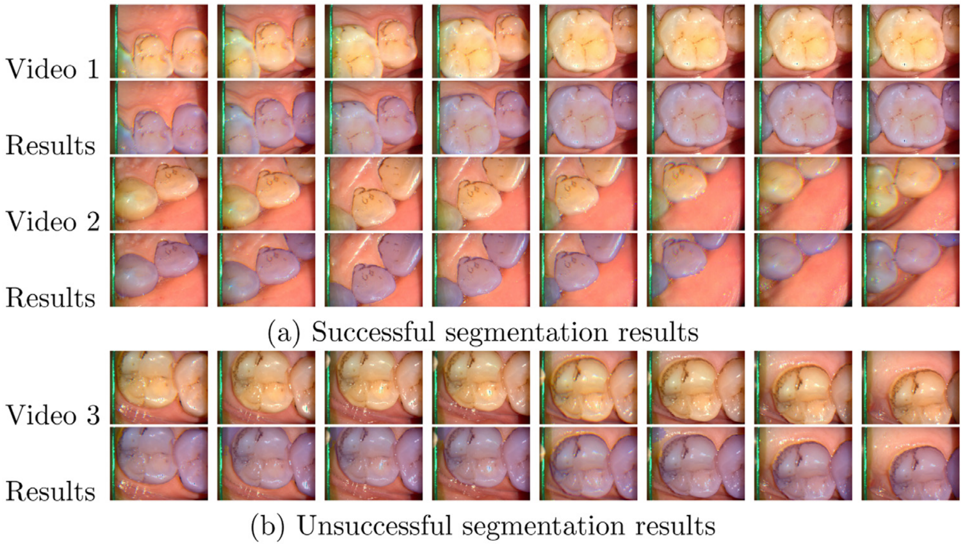

Video 1
Results
Video 2
Results
(a) Successful segmentation results
Video 3
Results
(b) Unsuccessful segmentation results

Fig. 9. Experimental results on the Shining3D dental data set. (a) Successful segmentation results. The first and third rows are the input video in sequence 1 and sequence 2, and the second and fourth rows are our segmentation results. (b) Unsuccessful segmentation results with a mistake on the top left of the image..

data set. Our method is a kind of mask propagation like PReMVOS that jointly learns and models spatial and temporal inference in a general probability formula. Our model obtains the best accuracy of all temporal context exploitation approaches.

The segmentation results of our approach are shown in Fig. 8. The upper row is the video input (frame numbers 0, 4, 8, 12, 16 and 20 in the bear sequence). The middle row (in red) shows our segmentation results, and the lower row (in white) shows the corresponding ground truth.

Our approach is consistently accurate in output object, but some object details may be unclear, for example, the foot of the bear in Fig. 8. The inexact segmentation of details is partially due to feature learning in multiscale analysis and partially due to non-adaptive threshold selection in temporal context distribution. What happens is that the warped feature is controlled by optical flow in the front layer, which lacks semantic and fine detailed information. Hence, small deviations in the front layer are propagated and enlarged in follow-up operations.

The posterior distribution of optical flow is affected by noise factors. Determining how to accurately model the noise factors is important in real applications. In our experiment, we use grid searching to set the Gaussian parameters. Other search approaches such as Bayesian optimization can be employed for parameter tuning.

4.7. Evaluation on the Shining3D dental data set

In this section, we discuss our approach compared to other approaches based on performance on the Shining3D dental data set. The results are shown in Table 1, with some image output results depicted in Fig. 9.

Note that our results on Shining3D give conclusions similar to those from our results on the DAVIS16 data set. The warped features or masks yield comparable Jaccard index accuracy. Moreover, our learning-based optical flow inference performs better than approaches that learn the mapping directly from successive frames.

Some of our actual results on the Shining3D dental data set are shown in Fig. 9. The successful segmentation results in Fig. 9(a) demonstrate that our method is robust to variation in dental shape, camera motion, and background clutter. Usually, lighting conditions are poor in mouth images, and such conditions will lessen the quality of the algorithms results. For this reason, we corrected the input images for brightness.

Fig. 9 (b) shows the unsuccessful segmentation results. Notice that a region on the top left of the image is responsible for some incorrect results, partially because boundaries may fade due to brightness correction. On the one hand, the region of shadow on the left and right image boundary increases the intra-variance of the teeth and other soft tissues, making the segmentator handle the multimodel sample poorly; on the other hand, the vignetting effect increases the diversity of the samples, and the performance deteriorates if only the center region, not the whole outline of the images, is selected for training and testing.

Objects in scenes with irregular lighting, or those that are occluded or truncated, are not well segmented by any method. In the future, we will extend our method to handle situations in which lighting is uneven, the perspective is irregular or the image is blurry. The processing time for each 640 × 480 image in

Shining3D dental data set is approximate 45 milliseconds, and we will also work on ways to increase efficiency in video segmentation so that our method will be practical to use for real-time applications such as autonomous driving.

5. Conclusion

Our experimental results show that active learning is an effective approach to video segmentation. Our framework works better for segmentation than do random-based approaches to sampling. Specifically, we present a method for probability sampling to infer temporal context and select the samples in a video sequence that are most representative and informative. Even though our method is weak in that it may arrive at inexact segmentation due in part to feature learning in multiscale analysis and nonadaptive threshold selection in temporal context distributions, our method is robust to variation in object shape, camera motion, and background clutter. Future work on the video segmentation problem could involve handling blurry objects and uneven lighting, as well as speeding up the segmentation process.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1403200, in part by the National Natural Science Foundation of China under Grant 61602407, Grant 61972351, and Grant 61572436, in part by the Scientific Research Foundation of National Health and Family Planning Commission under Grant WKJ-ZJ-1814, in part by the Key R&D Program of Zhejiang Province under Grant 2019C01004 and Grant 2019C03002, in part by the Natural Science Foundation of Zhejiang Province under Grant LY19F030005 and Grant LY18F020008, in part by the Hangzhou Major Science and Technology Innovation Project under Grant 20172011A038.

References

[1] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach.Intelligence 41 (7) (2018) 1531–1544.

[2] W. Wang, J. Shen, F. Porikli, R. Yang, Semi-supervised video object segmentation with super-trajectories, IEEE Trans. Pattern Anal. Mach.Intelligence 41 (4) (2019) 985–998.

[3] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, L. Shao, Real-time superpixel segmentation by dbscan clustering algorithm, IEEE Trans. Image Process. 25 (12) (2016) 5933–5942.

[4] Y. Tian, J. Gelernter, X. Wang, J. Li, Y. Yu, Traffic sign detection using a multi-scale recurrent attention network, IEEE Trans. Intell. Transp.Syst. (2019) 1–10.

[5] X. Lin, J.R.C. Pla, M.P. Feliu, One shot learning for generic instance segmentation in rgbd videos, in: in: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2019, pp. 233–239.

[6] G.F. Campos, S. Barbon, R.G. Mantovani, A meta-learning approach for recommendation of image segmentation algorithms, in: in: SIBGRAPI Conference on Graphics, Patterns and Images, 2016, pp. 370–377.

[7] J. Shen, J. Peng, L. Shao, Submodular trajectories for better motion segmentation in videos, IEEE Trans. Image Process. 27 (6) (2018) 2688–2700.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 801–818.

[9] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, A. Sorkine-Hornung, Learning video object segmentation from static images, in: in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2663–2672.

[10] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: in: Proceedings of the International Conference on Machine Learning, 2017, pp. 1183–1192.

[11] H. Song, W. Wang, S. Zhao, J. Shen, K.M. Lam, Pyramid dilated deeper convlstm for video salient object detection, in: in: Proceedings of the European Conference on Computer Vision, 2018, pp. 715–731.

[12] P. Tokmakov, C. Schmid, K. Alahari, Learning to segment moving objects, Int. J. Comput. Vis. 127 (3) (2019) 282–301.

[13] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S.C. Hoi, H. Ling, Learning unsupervised video object segmentation through visual attention, in: in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3064–3074.

[14] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: Unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3623–3632.

[15] L. Bao, B. Wu, W. Liu, Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf, in: in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5977–5986.

[16] P. Voigtlaender, B. Leibe, Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation, in: in: Proceedings of the British Machine Vision Conference, 2017, pp. 417–424.

[17] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L.V. Gool, Video object segmentation without temporal information, IEEE Trans. Pattern Anal. Mach.Intell. 41 (6) (2018) 1515–1530.

[18] J. Luiten, P. Voigtlaender, B. Leibe, Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018, in: In: The 2018 DAVIS Challenge on Video Object Segmentation-CVPR Workshops, 2018, pp. 6–14.

[19] Z. Wang, J. Xu, L. Liu, F. Zhu, L. Shao, Ranet: Ranking attention network for fast video object segmentation, in: in: Proceedings of the International Conference on Computer Vision, 2019. P. in press

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[21] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: in: Proceedings of the International Conference on Machine Learning, 2016, pp. 1050–1059.

[22] R. Wang, X.-Z. Wang, S. Kwong, C. Xu, Incorporating diversity and informativeness in multiple-instance active learning, IEEE Trans. Fuzzy Syst. 25 (6) (2017) 1460–1475.

[23] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, D. Tao, Exploring representativeness and informativeness for active learning, IEEE Trans. Cybern. 47 (1) (2017) 14–26.

[24] S. Jin, A. RoyChowdhury, H. Jiang, A. Singh, A. Prasad, D. Chakraborty, E. Learned-Miller, Unsupervised hard example mining from videos for improved object detection, in: in: Proceedings of the European Conference on Computer Vision, 2018, pp. 307–324.

[25] X. Li, F. Yang, H. Cheng, W. Liu, D. Shen, Contour knowledge transfer for salient object detection, in: in: Proceedings of the European Conference on Computer Vision, 2018, pp. 355–370.

[26] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, J. Wang, Deepsaliency: multi-task deep neural network model for salient object detection, IEEE Trans. Image Process. 25 (8) (2016) 3919–3930.

[27] F. Yang, X. Li, H. Cheng, Y. Guo, L. Chen, J. Li, Multi-scale bidirectional fcn for object skeleton extraction, in: in: AAAI Conference on Artificial Intelligence, 2018, pp. 420–427.

[28] X. Li, F. Yang, H. Cheng, J. Chen, Y. Guo, L. Chen, Multi-scale cascade network for salient object detection, in: in: Proceedings of the ACM international conference on Multimedia, 2017, pp. 439–447.

[29] X. Li, F. Yang, L. Chen, H. Cai, Saliency transfer: an example-based method for salient object detection, in: In: International Joint Conferences on Artificial Intelligence, 2016, pp. 3411–3417.

[30] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Process. 27 (1) (2017) 38–49.

[31] H. Lei, T. Han, F. Zhou, Z. Yu, J. Qin, A. Elazab, B. Lei, A deeply supervised residual network for hep-2 cell classification via cross-modal transfer learning, Pattern Recognit. 79 (2018) 290–302.

[32] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Models matter, so does training: an empirical study of cnns for optical flow estimation, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, pp. 201–208.

[33] V. Nekrasov, C. Shen, I. Reid, Light-weight refinenet for real-time semantic segmentation, in: Proceedings of the British Machine Vision Conference, 2018, pp. 278–284.

[34] D. Li, Y. Chen, M. Gao, S. Jiang, C. Huang, Multimodal gesture recognition using densely connected convolution and BLSTM, in: Proceedings of the International Conference on Pattern Recognition, 2018, pp. 3365–3370.

[35] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow, in: Symposium on Operating Systems Design and Implementation, 2016, pp. 265–283.

[36] F. Perazzi, J. Pont-Tuset, B. McWilliams, L.V. Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 724–732.

[37] Shining3d dental data set, 2019, https://drive.google.com/drive/folders/1Jds1LFVK0xuCJ7X9MU0iu97fmeIWuFLu?usp=sharing.

[38] Q. Zhou, X. Liang, K. Gong, L. Lin, Adaptive temporal encoding network for video instance-level human parsing, in: Proceedings of the ACM International Conference on Multimedia, 2018, pp. 1527–1535.

[39] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, T. Huang, Youtube-vos: Sequence-to-sequence video object segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 585–601.

[40] S.W. Oh, J.-Y. Lee, K. Sunkavalli, S.J. Kim, Fast video object segmentation by reference-guided mask propagation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7376–7385.

[41] H. Xiao, J. Feng, G. Lin, Y. Liu, M. Zhang, Monet: Deep motion exploitation for video object segmentation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 1140–1148.

[42] H. Lee, G. Park, H. Kim, Effective integration of morphological analysis and named entity recognition based on a recurrent neural network, Pattern Recognit. Lett. 112 (2018) 361–365.

[43] Z. Pei, X. Qi, Y. Zhang, M. Ma, Y.H. Yang, Human trajectory prediction in crowded scene using social-affinity long short-term memory, Pattern Recognit. 93 (2019) 273–282.

Yan Tian received Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2011. Then he had a postdoctoral research fellow position (2012–2015) in Zhejiang University, Hangzhou, China. He is currently an Associate Professor in the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. His current interests are machine learning and computer vision.

Guohua Cheng is Ph.D. candidate in Fudan University, Shanghai, China, and he received Master's degree from Nanyang University of technology, Singapore. He is currently 2012-Present, CEO of Jianpei Technology Co.,Ltd, 1000 Talents Plan member of Zhejiang province, 521 Program member of Hangzhou city, chairman of the Organizing Committee of the West Lake International Medical Forum, director of Digital China Industry Development Alliance, deputy leader of Big Data and Artificial Intelligence Group of China Research Hospital Association, director of the Artificial Intelligence Committee of China Association for Medical Device Industry, director of the Chinese Innovative Alliance of Industry Education, Research and Application of Artificial Intelligence for Medical Imaging (CAIERA), vice president of the Artificial Intelligence Alumni Association of Shanghai Jiaotong University, and executive director of Hangzhou Association for Artificial Intelligence. His current interests are machine learning and biomedical engineering, and he also works on medical image artificial intelligence.

Judith Gelernter received her Ph.D. in information science from Rutgers University in 2008. She did research from 2008 to 2015 in the Language Technologies Institute of the School of Computer Science of Carnegie Mellon University. She went on to become a Research Scientist in the Information Technology Laboratory at the National Institute of Standards and Technology (NIST) from 2015 to 2018. Presently, she is again affiliated with Rutgers University.

ShiHao Yu was born in Jinhua, China. Now he conducts as a research assistant in school of Computer Science and Information Engineering, Zhejiang Gongshang University, China. His current interests are deep learning and pattern recognition, and he also works on image retrieval.

Chao Song received Ph.D. degree from Zhejiang University, Zhagnzhou, China, in 2009. He is an Associate Professor in the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. His research interests are in virtual reality, mobile graphics, data mining and mobile game.

Bailin Yang received Ph.D. degree from Zhejiang University, Zhagnzhou, China, in 2007. He is a Professor in the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. His research interests are in virtual reality, mobile graphics, data mining and mobile game.