Measurement 194 (2022) 111003

Contents lists available at ScienceDirect

Measurement

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

An automatic quality evaluator for video object segmentation masks

Jingchun Cheng, Jiajie Song, Rui Xiong *, Xiong Pan, Chunxi Zhang

Beihang University, China

ARTICLE INFO

Keywords:
Mask quality estimation
Video object segmentation
Objective quality prediction
Deep learning

ABSTRACT

Video object segmentation (VOS) has been a research hot-spot these years. However, evaluating the performance of different VOS methods requires labor-intensive and time-consuming manually labeled mask annotations, making it hard to validate the algorithm quality in field tests. In this paper, we tackle the problem of automatically measuring the mask quality for video object segmentation tasks without accessing manual annotations. We propose that with an elaborately designed network structure, we can extract quality-sensitive features to predict mask quality scores without ground-truth labels. To achieve this, we train an end-to-end convolutional neural network to capture the quality-sensitive features with both spatial reference and temporal reference. In the proposed Video Object Segmentation Evaluation Network, the *VOSE-Net*, the corresponding video frame and motion amplitude information are used for spatial and temporal references respectively. Instead of directly concatenating features for mask and references, we extract spatial quality cues with feature correlation, which is more rational and effective in this specific task. Taking in the segmented mask, its corresponding frame image and optical flow map, the *VOSE-Net* can provide an accurate quality estimation without the need for human intervention. To train and verify the proposed network, we construct a new dataset by using the DAVIS video segmentation benchmark and results from many public video object segmentation algorithms. We also demonstrate the robustness and usefulness of the proposed method on several applications, i.e. proposal selection, parameter optimization, arbitrary video mask evaluation. The experimental results and analysis show that the *VOSE-Net* is fast, effective and of practical use.

1. Introduction

Following the increasing demand for video analysis, the primary video object segmentation (P-VOS) task has become a research focus [1–5]. It requires methods to track and segment the most dominant instance in videos, and is also called single-instance video object segmentation, unsupervised video object segmentation, etc. Numerous methods have been developed in this field for the past decades [6,7], producing more and more accurate target masks [8], and segmenting with less and less processing time [9,10]. However, current algorithms heavily rely on manual annotations to train and validate their models, e.g. using the densely-annotated datasets [3]. Video object segmentation methods need the ground-truth annotations for test videos to compute the overall mask quality with evaluation metrics like the Jaccard similarity, F-score [11] (we show some examples of mask prediction and their computed Jaccard scores in Fig. 1). This limits the progress of P-VOS, for that the ground-truth labels are hard to get in practical applications and the methods cannot obtain the objective evaluation scores to assess their ability or adjust parameters online.

Based on this observation, we aim to provide a blind quality measurement algorithm for primary video segmentation quality assessment (VSQA), which can automatically estimate the mask quality scores without the need of human annotations. This technique is known as segmentation quality assessment (SQA) when applied on images [12]. Researchers have come up with several SQA methods [12–14], the common idea of which is to extract cues from the segmentation mask with its corresponding original image, and then regress the features into the quality score. As SQA is a challenging problem, most methods limit the task to patch-wise or bounding box restrained regions where each image patch contains a single instance, or add in weak human supervision [15] during prediction. Among the multiple SQA methods, deep learning based models [12,14] show superior ability, similar to what happens in other computer vision fields (image classification, object detection, image segmentation, etc.). This is because of the strong learning ability and powerful fitting capacity of deep neural networks. Take [12] for an example, by feeding the original image and segmentation patches into a deep neural network, [12] successfully obtains an image segmentation evaluator, demonstrating that CNN is

* Correspondence to: Institute of Optics and Electronics, Beihang University, Beijing, China.
E-mail addresses: chengjingchun14@163.com (J. Cheng), songjiajie@buaa.edu.cn (J. Song), ruix2017@buaa.edu.cn (R. Xiong), 08768@buaa.edu.cn (X. Pan), zhangchunxi@buaa.edu.cn (C. Zhang).

https://doi.org/10.1016/j.measurement.2022.111003
Received 29 November 2021; Received in revised form 9 February 2022; Accepted 5 March 2022
Available online 19 March 2022
0263-2241/© 2022 Elsevier Ltd. All rights reserved.

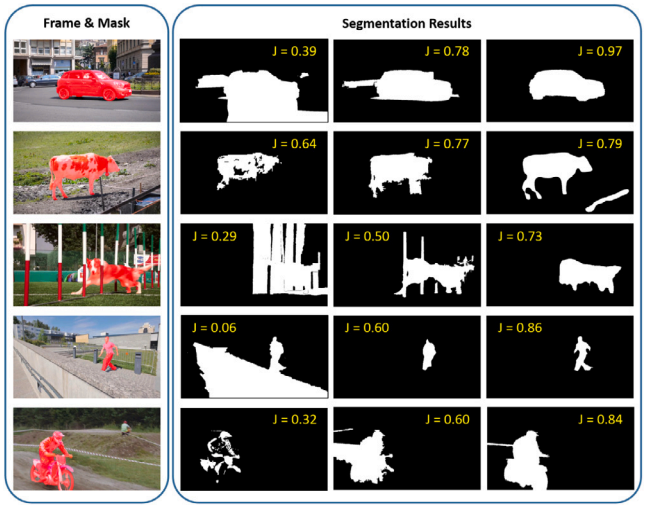

Fig. 1. Illustration of masks with various qualities from the proposed VISA dataset. The left column gives some examples of the video frames with their pixel-wise manually annotated masks (colored in red). Several segmentation results of different qualities are shown on the right, where J denote the Jaccard similarity scores (an objective score evaluating the matching rate between mask and the ground-truth label, also known as IoU).

capable of depicting mask-image correspondence which is adequate for mask quality prediction.

In this paper, we extend the image segmentation quality assessment to the video level. Different from images, masks from adjacent video frames are not independent. They are closely related in the temporal space (connected with object and background motions). Therefore, in the proposed evaluator, we incorporate both spatial and temporal features for video mask estimation. We build an end-to-end convolutional neural network, the *VOSE-Net* (Video Object Segmentation Evaluation Net), to estimate the video mask quality scores. The *VOSE-Net* employs the original frame image for its spatial reference, and the motion map (optical flow) for the temporal reference. Instead of crudely concatenating features from references and segmentation mask, the *VOSE-Net* fuses the spatial correlations with temporal information, and maps into one comprehensive score for the input mask. The overall framework has an end-to-end structure like shown in Fig. 2, where mask quality of a test video frame can be obtained via a single forward at the speed of about 70 ms.

To train and validate the proposed VOSE-Net, we construct a new dataset from the densely-annotated video object segmentation dataset [3] and various public video segmentation models. The Jaccard similarity scores between mask-frame pairs are used as ground-truth scores for evaluating the network ability. Furthermore, we compare and analyze different feature fusion methods for segmentation quality evaluation and experiment on three application scenarios (i.e. segmentation proposal selection, post-processing parameter optimization, quality evaluation on arbitrary videos) in Section 5.2. We show that the proposed *VOSE-Net* is robust, generally applicable, and of practical use.

Overall, our main contributions are as follows:

(1) we propose a deep mask quality assessment network for primary video object segmentation;
(2) we construct a new dataset for objective video mask quality estimation task, which can encourage the development of researches in VSQA;
(3) we validate that the proposed *VOSE-Net* has satisfying quality assessment ability and can be applied to a variety of related tasks.

2. Related work

2.1. Video object segmentation

Video object segmentation aims at separating target object from background at pixel level. Depending on whether an initial annotation frame is given in each test video, it can be classified into 'unsupervised' video object segmentation and 'semi-supervised' video object segmentation. 'Semi-supervised' video object segmentation methods are considered to have known a manual object mask for the target instance(s) in the first frame of videos. They have a specific concept of the target instances to help them track and segment objects throughout videos. Test videos with 'semi-supervised' setting can have more than one target instance, as each of them can be manually defined in the initial frame. 'Unsupervised' video object segmentation, also known as primary video object segmentation, has no prior knowledge of the target instance. When testing upon a new video, P-VOS algorithms do not need manual first frame annotations, they are able to automatically judge and locate onto the dominant video object, then separate it from background pixels. P-VOS methods can be more easily applied to practical uses because they have no burden for manual labels. As no prior knowledge is given, test videos are required to have predominant instances. Our proposed evaluator is designed for primary video object segmentation, where the test videos are assumed to each have a predominant target instance that algorithms can discover and segment out.

A large number of video object segmentation techniques have been developed recent years [8–10,16–23]. These methods are based on graph models [19], object appearance [16,24], similarity cues [10,25], etc. Some common informative cues they count on for video segmentation are the appearance cue, motion cue, memory cue, etc. For example, [16] deals with video segmentation in a frame-independent order, it over-fits a general foreground segmentation network on each test video via heavily fine-tuning on the initial mask; [7] aggregates the appearance cue with motion features to find an optimal feature combination in video segmentation; [23] constructs the space–time correspondences of video context for efficient video object segmentation; and [25] exploits the collaborative pixel-level matching with instance-level attentions to generate accurate mask predictions. The common motion cue, optical flow, is often used to measure the pixel movement correspondence in adjacent frames; while similarity cues might be used to measure the appearance correspondence across a longer video clip (e.g. the re-identification network used in [26]). In this paper, we propose to use both the appearance cue and the motion cue for evaluating video masks, i.e. the spatial reference and the temporal reference.

2.2. Segmentation quality evaluation

Segmentation quality assessment task is very different from the long and widely studied image quality assessment task. Image quality assessment (IQA) aims to estimate the quality of distorted images which endure transmission loss, compression loss, etc. Even with the original full-quality image as a reference, the quality score of a distorted image is not quantifiable; therefore, the ground-truth scores for image quality assessment datasets [27–30] are usually obtained by manual definition or averaged human subjective scores.

In contrast, segmentation quality assessment has several universally-accepted objective evaluation scores (e.g. the Jaccard similarity) given the ground-truth object mask as reference. Algorithms for segmentation quality assessment target at estimating the quality of a segmentation mask. They can be roughly categorized into 'non-blind' and 'blind', i.e. with or without accessing the manual segmentation labels. The 'non-blind' methods know both the to-be-evaluated segmentation mask and the annotation; they manage to measure the correspondence between the test mask and the ground-truth annotation with mathematical analysis. For example, the commonly used segmentation evaluation

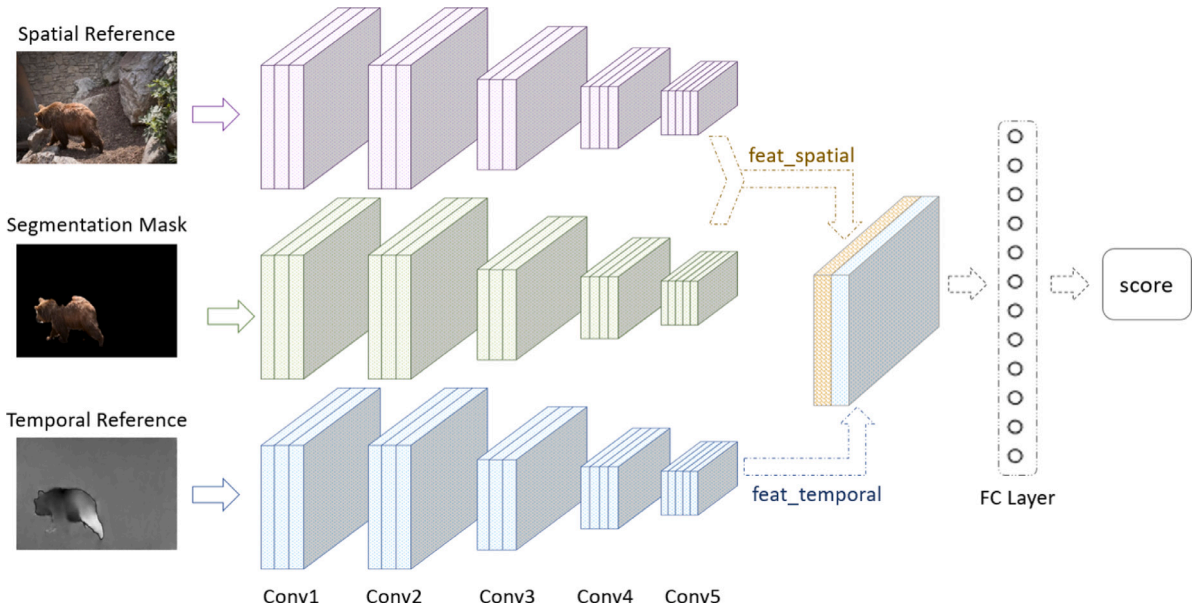

Fig. 2. Framework of the *VOSE-Net*. This figure shows the overall framework for the proposed quality evaluator, the *VOSE-Net*. Guided by the spatial correlation reference and the temporal motion information, the network can automatically estimate the quality score (0–1) of a video segmentation mask. Note that both references share the same convolution body net with the mask input (Conv1–Conv5).

metric IoU [11] (also known as Jaccard similarity) computes the intersection over union between ground-truth masks and predicted segmentation; and the distance functions such as Hausdorff distance, mean distance [31] and F1 measurement [32] are used to evaluate the preciseness of segmentation border pixels (boundary accuracy). These criteria are widely used in image segmentation [33], video segmentation [3] and foreground segmentation tasks. For 'blind' estimation, algorithms are required to assess the quality of a segmentation mask with the absence of pixel-wise mask annotations. Seeing the rise of CNN and its outperforming achievements in computer vision tasks, methods [12–14] incorporate deep neural network for this task. As there are many challenging factors (unknown object category, large instance deformation, masks with various types of flaw), most of these evaluation methods restrain the candidate area to a bounding box which tightly surrounds a single instance. They feed segmented image patch and the corresponding unprocessed patch to different structured networks, train the network with objective quality scores (e.g. IoU), and then enable the network to predict scores for test images without manual annotations.

Most related to our method is [12], where the author uses a VGG-16 [34] network to predict segmentation quality from segmented patches located in object bounding boxes. Different from [12], we extend the segmentation mask quality evaluation to the video domain by adding in unique video cues with a more delicate network design, carry out extensive comparisons and analysis for network feature fusion strategies, as well as present several practical application scenarios.

3. The proposed method

3.1. The VOSE-Net

Inspired by the effectiveness of residual networks in computer vision tasks [35–37], we build our *VOSE-Net* from the widely-used ResNet-101 [4,38,39] structure. There are five main convolution blocks (*Conv1* to *Conv5*) in the ResNet-101, outputting feature maps with sizes of 1/4, 1/4, 1/8, 1/16 and 1/32 of the input size respectively (see Fig. 2). As demonstrated in [40], with the network going deeper, the extracted feature condenses and promotes to a higher semantic level.

In that videos are sequences of consecutive frames whose information lie in both spatial and temporal domains, we incorporate both spatial-domain and temporal-domain references for the mask quality evaluator. As illustrated in the framework of the *VOSE-Net* (Fig. 2), we take the convolutional features from *Conv5* with the highest semantic level and the largest receptive field as our descriptions for the segmentation mask, spatial reference and temporal reference, which we use f_m, r_s and r_t to represent in the following. After extracting these feature descriptions, we combine the spatial-domain information with the temporal-domain information, and feed the aggregated feature into a fully connected layer for predicting a specific quality score for the input segmentation mask.

For the spatial information, we take the corresponding video frame as the reference, which is similar to [12]. [12] directly concatenates or element-wisely sums the spatial-reference feature (r_s) and the mask feature (f_m), and proves that these simple operations can provide enough quality information for image segmentation evaluation. In this paper, we propose that their correlation information should help the evaluation process more as the reference image and segmentation mask inputs are closely related. We verify this point by experiments in Section 5.1. In the proposed framework, we process r_s and f_m with a correlation layer to obtain the spatial-domain information.

The correlation layer is designed to perform multiplicative patch comparisons between two feature maps by [41]. For the feature maps of spatial reference (r_s) and segmentation mask (f_m), the correlation value is computed as:

$$\mathbb{C}(\vec{p}_r, \vec{p}_f) = \sum_{\vec{o}\in[-r,r]\times[-r,r]} r_s(\vec{p}_r + \vec{o}) \times f_m(\vec{p}_f + \vec{o}). \tag{1}$$

In Eq. (1), $\vec{p}_r$ and $\vec{p}_f$ denote the position coordinates on feature maps r_s and f_m respectively. Let $W \times H \times C$ be the width, height, and channel number of r_s and f_m, then there is: $\vec{p}_r, \vec{p}_f \in \{(w,h)|0 \le w < W, 0 \le h < H\}$. r determines the scope of candidate patch whose square edge length is $2r+1$, and the correlation operation $\mathbb{C}(\vec{p}_r, \vec{p}_f)$ computes the relationship between patch centered at $\vec{p}_r$ on r_s and patch centered at $\vec{p}_f$ on f_m. For simplicity, the correlation is only carried out within a local range where $\vec{p}_r - \vec{p}_f \in [-d,d]\times[-d,d]$. We pick the optical range value of d through experimental exploration in Section 5.1.

In the proposed *VOSE-Net*, we set the correlation scope to be specific to point-wise ($r = 0$), and the local computation range to be a rather large value ($d = 4$ compared with the feature map size $W = 10, H = 16$). In this case, the computation of $\mathbb{C}(\vec{p}_r, \vec{p}_f)$ involves $C * (2d+1)^2$

multiplications. As the correlation layer directly convolves feature map with feature map, there are no trainable weights. This layer takes in the same-size feature maps r_s and f_m, and outputs their correlation values with the size of $W * H * (2d+1)^2$ (in practical implementation, the relative displacements of $(2d+1)^2$ are organized in channel, so the output is 3-dimensional instead of 4).

For the temporal information, we employ the commonly-used optical flow [42–44] to represent the inter-frame time–space relationship. We propose that pixels in the same instance tend to have similar movement between adjacent frames. For effectiveness and efficiency, we choose a state-of-the-art, CNN-based optical flow algorithm, the PWC-Net [44] to generate our temporal reference. We also compare with two other optical flow algorithms [42,43] in Section 5.1 to show the impact of optical flow on the overall performance. In practice, we calculate the motion amplitude map which reflects the degree of motion as the temporal-domain reference. The extracted temporal feature with dimension of 2048 is decreased to the dimension of 64 by 1×1 convolution filters, in order to prevent spatial information (dimension of 81) from being overwhelmed by the temporal feature. These two features are then concatenated for the final quality prediction process.

For simplicity, the three inputs (spatial reference, temporal reference, segmentation mask) share the same parameters from *Conv1* to *Conv5*. Note that to do this, the temporal-domain reference image is duplicated to three channels. After the fully-connected layer, the output goes through a Sigmoid function to be normalized within $[0, 1]$, then the final estimation score can be expressed as:

$$score = \frac{1}{1 + e^{-(\mathbb{F}([\mathtt{info}_s, \mathtt{info}_t]; W))}}, \tag{2}$$

where $\mathbb{F}(.)$ denotes the fully-connected layer with weights W, the spatial information and temporal information $\mathtt{info}_s$ and $\mathtt{info}_t$ are the spatial correlation output $\mathbb{C}(r_s, f_m)$ and motion amplitude descriptions r_t respectively.

3.2. Training details

To train and validate the proposed method, we construct a new video mask quality evaluation database named the VIdeo Segmentation Assessment (VISA) dataset. It provides us with the video frame, segmentation masks and the corresponding objective quality scores (Jaccard similarities). The collection and detailed settings of the VISA dataset will be introduced in Section 4.1.

During training, we use the L1 Loss which is defined as:

$$L_1(X_i) = |\mathcal{N}(X_i) - y_i|, \tag{3}$$

where $\mathcal{N}(.)$ denotes our quality estimation network; X_i and y_i denote the segmentation input and its corresponding ground-truth score, respectively.

Standard Stochastic Gradient Descent (SGD) is used for optimizing the proposed *VOSE-Net*, with the learning rate set to 1e-4, momentum of 0.9, and batch size of 20 (due to memory limitation, we adjust the value of average_loss to set the batch size). The initial learning rate is decreased by 0.1 every 50'000 iterations for a total of 100'000 iterations. This whole process is carried out on a single Titan X GPU, and it takes about 12 h to get a convergent model.

3.3. Applications

As mentioned above, a blind mask evaluator is very useful in practical applications. It can be used to select the most likely object mask among hundreds of segmentation proposals, or help search the best set of parameters for post-processing methods. Most importantly, this evaluator can help estimate the performance of different algorithms on raw videos like Internet videos, amateur videos, homemade videos, etc. Very different from the ones in the video segmentation datasets where methods are trained upon, raw videos face the conundrum of choosing the best-performing segmentation algorithm without human intervention. We propose that this can be solved by our automatic quality evaluator, the *VOSE-Net*. We demonstrate in Section 5.1 that with the *VOSE-Net*, we are able to estimate the general ability of various unsupervised video segmentation methods and choose the proper one for each specific test video. Detailed analysis, quantitative comparisons and visual illustrations are shown in Section 5.2.

4. Dataset construction

We construct a new dataset of video frames and segmentation masks from various qualities for training and testing video mask quality evaluators, the VIdeo Segmentation Assessment (VISA) dataset.

4.1. Data and distribution

We collect data from the DAVIS 2016 dataset [3], with all the video frame images and the segmentation masks from both semi-supervised [4,5,8–10,16–20,24,45–54] and unsupervised [6,7,55–60, 60–65] algorithms in order to cover a wide range of mask qualities. In total, we obtain about ninety thousand segmentation masks with their original frames. The ground-truth label for each segmentation mask is set to be the Jaccard similarity score [11] (also referred to as the IoU, intersection-over-union) value:

$$J(I_m, I_{gt}) = \frac{|I_m \cap I_{gt}|}{|I_m \cup I_g t|} = \frac{|TP|}{|TP| + |FN| + |FP|}, \tag{4}$$

where I_m and I_{gt} represent the segmentation mask and ground-truth annotation respectively; TP, FN and FP denote the true positive, false negative and false positive pixel points, respectively; $|.|$ counts the number of pixels within each set. Among the three commonly used evaluation metrics, i.e. the Jaccard similarity (J), boundary score (F) and temporal instability (T), we choose J to represent the overall mask quality for that J can better illustrate the overall qualities of segmentation masks, and that higher J scores often accompany with higher F scores.

The segmentation masks are split into a training set of 66'474 images and a testing set of 12'396 images, where the videos do not overlap in the training and testing set (i.e. masks for 10 categories only appear in the testing set, while masks for the other 40 categories only appear in the training set). To make the evaluation results on this dataset more convincing, we also put masks from three algorithms only in the testing set so that the learning methods cannot obtain their segmentation pattern via training.

Fig. 1 shows some examples of frame images, manual annotations, segmentation masks and ground-truth scores (J) from the VISA dataset. We draw the mask quality distribution in the VISA dataset in Fig. 3, where we can see that the masks within this database cover a wide range of qualities, including error segmentations (masks that do not overlap with ground-truth labels) and perfect predictions (masks that are exactly the same with ground-truths). To extend to a larger range of mask types, we also carry out a specific mask warping data augmentation process during training. Details for the mask warping technique are introduced in Section 4.3.

4.2. Evaluation criterion

We use four evaluation metrics to compare the ground-truth Jaccard similarity with the predicted score by the *VOSE-Net*, i.e. MAE, RMSE, PLCC and SRCC.

MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) are two commonly used criteria for evaluating the preciseness of variable predictions. MAE computes the average absolute error between each pair of predicted variable and its ground-truth value:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |J_i - pred_i|, \tag{5}$$

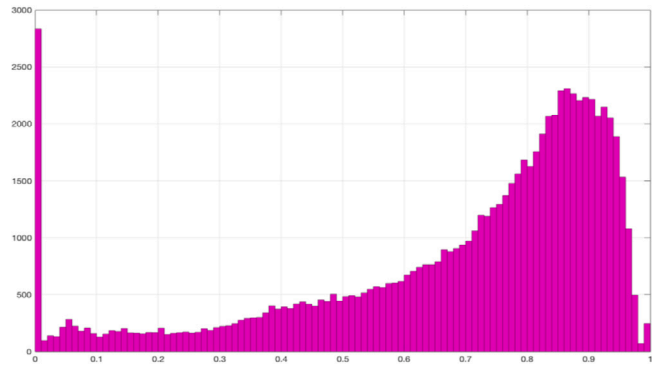

Fig. 3. Distribution of ground-truth Jaccard scores in the VISA dataset.

where N is the number of image set, J_i is the ground-truth Jaccard similarity (Eq. (4)), and $pred_i$ stands for the network prediction value. Similarly, RMSE computes the root value of squared error to measure the difference between prediction and labels:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(J_i - pred_i)^2}. \tag{6}$$

As MAE measures the mean values of differences between predicted quality scores and ground-truths and RMSE measures the standard deviation between predictions and ground-truths. The combination of these two metrics can be used to evaluate the effectiveness of the quality prediction models. For MAE and RMSE $\in [0, \infty)$, better prediction methods have lower values.

PLCC (Pearson Linear Correlation Coefficient) and SRCC (Spearman's Rank Correlation Coefficient) are two widely used metrics to measure the affinity between two groups of variables. The equations for PLCC and SRCC are defined as follows:

$$PLCC = \frac{\sum_{i=1}^{N}(J_i - \overline{J})(pred_i - \overline{pred})}{\sqrt{\sum_{i=1}^{N}(J_i - \overline{J})^2 \sum_{i=1}^{N}(pred_i - \overline{pred})^2}}, \tag{7}$$

$$SRCC = 1 - \frac{6\sum_{i=1}^{N}(R_{J_i} - R_{pred_i})^2}{N(N^2-1)}, \tag{8}$$

where $\overline{J}$ and $\overline{pred}$ are the averaged value of ground-truth set (J) and prediction set ($pred$); R_{J_i} and R_{pred_i} denote the ranks of values J_i and $pred_i$ respectively. Not like MAE and RMSE which measures accuracy for points independently, PLCC and SRCC evaluate the group distribution correlation in parametric and non-parametric ways. They both have the distribution between −1 to 1, where 0 denotes that the two point groups are not correlated; a higher positive value means the two groups are more positively correlated; and negative values illustrate negative correlations.

In this paper, we use MAE and RMSE to validate the accuracy of the *VOSE-Net* prediction, and the PLCC and SRCC to demonstrate that the *VOSE-Net* has a similar score distribution trend with manual labels (which means for better masks, the *VOSE-Net* has higher predicted scores, and vice versa).

4.3. Data augmentation

For train-time data augmentation, we apply two useful methods on the VISA dataset: the commonly-used affine transformation, and a self-designed mask warping algorithm.

Affine transformation. For affine transformation, like most of the video segmentation methods [4,16], we apply random flipping, rotation, scaling and translation operation to both masks and video frames. The transformation range is limited to: rotation angle $\in [-10, 10]$, scaling factor $\in [0.8, 1.5]$, and translation within 10% of the image width or height.

During training, each pair of adjacent video frames in the same training batch share the same set of transformation parameters so that their temporal connection stays unbroken.

Mask warping. To train with masks of more variety, we propose a mask warping augmentation method for segmentation masks. For mask warping, we randomly contaminate each input mask (not the video frame) with three different operations described as follows.

(1) *BLUR* operation randomly blurs the input mask with radius from 0 to 5 pixels to produce larger and worse masks.
(2) *HAIRY* operation adds horizontal and vertical burr to make the mask hairier, the percentage of edge burr is controlled below 30%.
(3) *DISTORTION* operation maps the mask pixels from one random quadrangle to another (position change within 16 pixels), resulting in a twisted target mask.

During training, each input mask is augmented 1000 times with a random combination of these three mask warping algorithms. We show some visual examples for these mask contamination operations in Fig. 4, where the ground-truth masks (*mask* in the second column) are contaminated with *BLUR*, *HAIRY* and *DISTORTION* operations respectively. To better illustrate how these operations influence mask quality, we also present their Jaccard similarity scores (Section 4.1) on the top-left corner of each converted mask. We can see that all three mask warping operations degrade the original mask quality with different types of noises, thus making the training data much richer.

Table 1
Ablation study on the VISA test set. This table shows the performance of the *VOSE-Net* without either reference (spatial, temporal) or training data augmentation steps (introduced in Section 4.3).

	MAE ↓	RMSE ↓	PLCC ↑	SRCC ↑
VOSE-Net	**0.034**	**0.052**	**0.980**	**0.958**
– spatial reference	0.106	0.155	0.805	0.761
– temporal reference	0.055	0.08	0.941	0.907
– affine transformation	0.04	0.063	0.970	0.946
– mask warping	0.039	0.06	0.973	0.951
– affine transformation – mask warping	0.047	0.075	0.958	0.938

5. Experimental results

In this experimental section, we demonstrate the accuracy, robustness and general applicability of the *VOSE-NET* via extensive experiments and analysis. Besides, we also carry out a thorough comparison of the network feature fusion, parameter settings; and show some practical applications. All our dataset, code and models will be made available to the public.

5.1. Automatic quality evaluation

In this part, we demonstrate the automatic mask quality estimation ability of the proposed *VOSE-Net*, as well as the effectiveness of its network structure, data augmentation process, and parameter settings. The networks are trained and tested on the VISA dataset's training set and testing set, respectively.

Ablation study

We conduct an ablation study to validate that each part of the proposed *VOSE-Net* has its own contribution to the overall performance. As shown in Table 1, we remove each separable part from the *VOSE-Net*, i.e. the spatial reference, temporal reference, data augmentation by affine transformation and by mask warping (defined in Section 4.3).

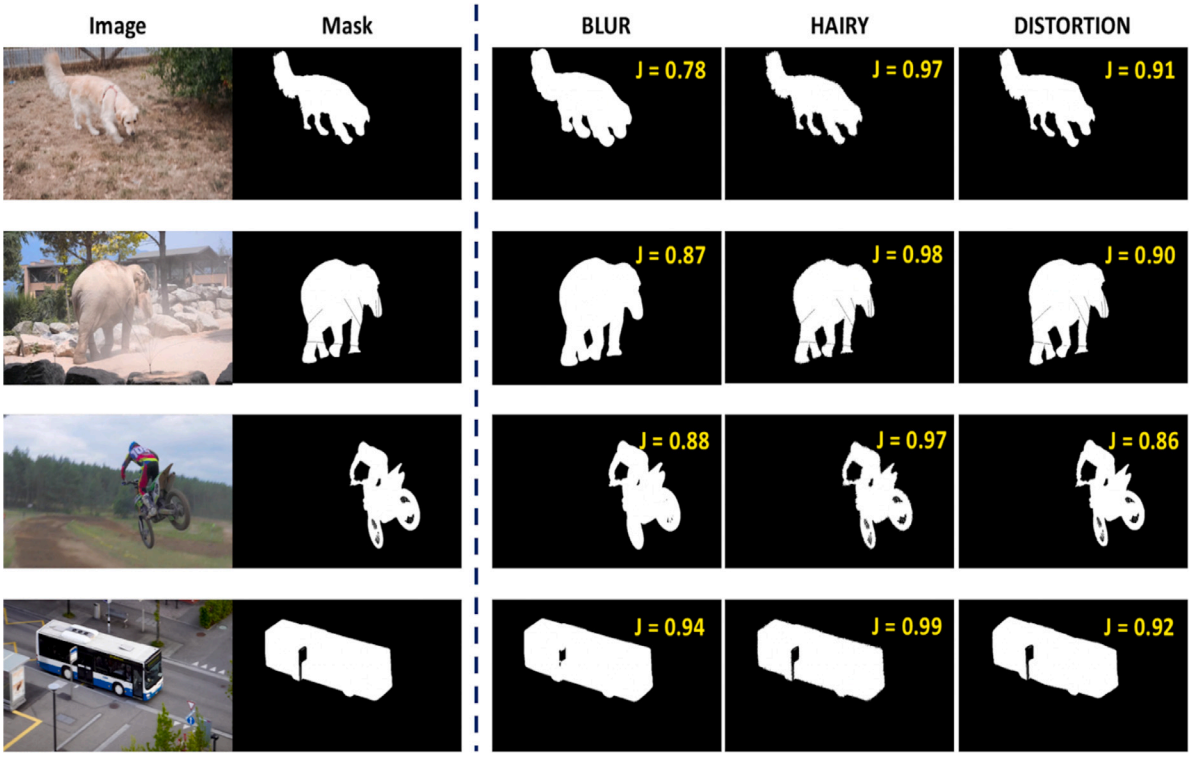

Fig. 4. Example results for mask warping operations in data augmentation. $Image$ and $Mask$ denote the original video image and the ground-truth annotation respectively. Columns 3 to 5 show the impact of mask warping operations (i.e. $BLUR$, $HAIRY$ and $DISTORTION$) when applied on the ground-truth mask. J illustrates the Jaccard score for each converted mask.

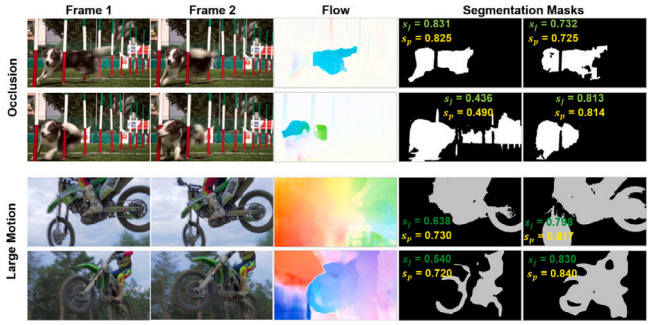

Fig. 5. Example results of the *VOSE-Net* predictions in cases of occlusion and large motion where the temporal reference flow is inaccurate. s_j and s_p denote the ground-truth score and the *VOSE-Net* prediction score, respectively.

We can see that for mask quality assessment, the spatial reference is of more importance than the temporal reference: removing spatial reference causes 0.051 more loss in MAE than removing the temporal reference. This is because the spatial reference (original frame image) can provide more detailed texture and color information than the temporal reference (optical flow): the comparison between a mask and its corresponding image can show the alignment of mask boundaries, image color piece distribution, mask completeness, etc. Besides, the temporal reference, optical flow, may fail to capture accurate temporal correlations between frames with occlusion or large motion like shown in Fig. 5. And in such cases, the spatial reference can keep providing accurate mask quality features for is it strongly correlated to the original image and is not affected by temporal variations.

We can also see that data augmentation plays an important role in the training process: without the affine data augmentation and the mask warping data augmentation, the overall MAE worsens from 0.034 to 0.04 (drops 17%) and 0.039 (drops 15%) respectively. This demonstrates that both augmentation methods can help enrich the data distribution and therefore make the network more robust and of higher prediction accuracy.

Table 2
Ability of the *VOSE-Net* to predict J, F and G scores.

	MAE ↓	RMSE ↓	PLCC ↑	SRCC ↑
VOSE-Net-J	0.034	0.052	0.980	0.958
VOSE-Net-F	0.061	0.084	0.945	0.938
VOSE-Net-G	0.043	0.65	0.966	0.951

Ability of predicting F and G scores

Although the *VOSE-Net* is designed to predict the overall performance (J scores) of video segmentation masks. We show that it is also capable of predicting the boundary accuracies (F scores) and the comprehensive G scores where $G = (J + F)/2$. We train the *VOSE-Net* on the VISA dataset with ground-truth scores for F and G respectively, and evaluate the performances in Table 2, where the *VOSE-Net* trained to predict J, F and G scores are denoted by $-J$, $-F$, $-G$. We can see that with the same training settings, the *VOSE-Net* predicts J scores most accurately, and is worst at directly predicting F scores. The reason that boundary accuracies are harder to be blindly estimated may be that the temporal references do not guarantee precise boundaries, and that less information is provided from the correlation between the original image and spatial reference on boundaries than the entire mask areas. This difficulty also lowers the precision in G score estimation, for it is a comprehensive score consisting of both J and F. We propose that as J score represents the intersection-over-union of the predicted mask and ground-truth mask, the *VOSE-Net* is able to predict the video segmentation qualities in most cases. For more accurate blind boundary quality estimation, further studies are still needed. In addition, the *VOSE-Net* cannot directly predict the temporal stability score T introduced in Section 4 for that this score computes

Table 3
Comparison of the *VOSE-Net* with different components on the VISA test set. Most of the networks are trained only with affine transformation for saving training time.

Settings	Input type	Max displacement	Optical flow	Data augmentation	MAE ↓	RMSE ↓	PLCC ↑	SRCC ↑
S1	M_{RGB}	1 (3)	PWC-Net [44]	Affine transformation	0.047	0.071	0.962	0.938
S2	M_{RGB}	2 (5)	PWC-Net [44]	Affine transformation	0.043	0.07	0.963	0.943
S3	M_{RGB}	3 (7)	PWC-Net [44]	Affine transformation	0.041	0.063	0.970	0.945
S4	M_{RGB}	4 (9)	PWC-Net [44]	Affine transformation	0.039	0.060	0.973	0.951
S5	M_{RGB}	5 (11)	PWC-Net [44]	Affine transformation	0.039	0.060	0.973	0.951
S6	M_{RGB}	6 (13)	PWC-Net [44]	Affine transformation	0.040	0.062	0.971	0.947
S7	M_{RGB}	7 (15)	PWC-Net [44]	Affine transformation	0.043	0.07	0.963	0.943
S8	M_{RGB}	8 (17)	PWC-Net [44]	Affine transformation	0.047	0.075	0.958	0.938
S4	M_{RGB}	4 (9)	PWC-Net [44]	Affine transformation	0.039	0.06	0.973	0.951
S9	M_{Bi}	4 (9)	PWC-Net [44]	Affine transformation	0.079	0.109	0.909	0.875
S4	M_{RGB}	4 (9)	PWC-Net [44]	Affine transformation	0.039	0.06	0.973	0.951
S10	M_{RGB}	4 (9)	EpicFlow [43]	Affine transformation	0.041	0.063	0.970	0.945
S11	M_{RGB}	4 (9)	FlowNet2[42]	Affine transformation	0.04	0.062	0.971	0.947
VOSE-Net	M_{RGB}	4 (9)	PWC-Net [44]	Affine transformation + Mask warping	0.034	0.052	0.980	0.958

the overall algorithm performance consistency over the videos, while the *VOSE-Net* measures short-term mask qualities.

Component analysis

The *VOSE-Net* has several variable components, i.e. the max displacement for correlation computation, the input mask form, and the optical flow computation methods.

The max displacement parameter in the correlation layer determines the range of correlation operation (see correlation definition in Eq. (1)) : if the max displacement is d pixel, then the computation range is $2d+1$ pixels. To obtain the optimal max displacement value, we search from the minimum value 1 to the size of the feature map ($S1-S8$ in Table 3), where we find that medium values 4 and 5 tie for the same best performance. To reduce computation, we choose 4 for the max displacement value in the *VOSE-Net*.

For the input mask form, we test two types of mask input: the binary mask prediction from video segmentation methods (M_{Bi}) and the colored mask M_{RGB} which is obtained by cutting the corresponding frame image with the binary mask. Comparing $S4$ and $S9$ in Table 3, it is easy to see that M_{RGB} makes a better input type. This accords with our conjecture, in that masks in the form of M_{RGB} keep more details and have better correspondence with their spatial references.

We also compare three state-of-the-art optical flow algorithms [42–44] to see how different optical flow qualities impact on the mask evaluation performance. Comparison of $S4$, $S10$ and $S11$ illustrates that the *VOSE-Net* is only minorly influenced by the changing of optical flow sources, which further validates the robustness of the *VOSE-Net*.

Based on the above comparisons, we choose the optimal component setting (Max Displacement: 4, Input Type: M_{RGB}, Optical Flow: PWC-Net) for the *VOSE-Net* as presented in the last row of Table 3.

Mask Quality evaluation with various feature fusions

The *VOSE-Net* integrates multiple knowledge representations [66] for predicting mask qualities, i.e. reinforcing between visual knowledge and deep representations, and then fusing deep representations to enable explicit reasoning of blindly estimating video mask qualities. Different from the commonly used late fusion for flow and RGB models, we delicately design the *VOSE-Net* structure to contrastively integrate RGB model features and complementarily add in flow model features for better prediction. In this part, we compare with several networks trained for different types of references and feature fusion schemes (including late fusion like *net*1, *net*3 and *net*7) to demonstrate the high performance and structure effectiveness of the *VOSE-Net*. All the networks are trained with the same train-set data of the VISA database (with the affine transformation data augmentation) and the same optimization parameters as described in Section 3.2. We use f_m, r_s and r_t to represent the features of segmentation mask, spatial and temporal references respectively; we use $\oplus$ to denote the feature concatenation operation, and $\otimes$ to denote the feature correlation operation. As described in Section 3.1, we use the original frame image as the spatial reference r_s, and the optical flow motion amplitude as the temporal reference r_t.

Table 4
Network structure comparison. Performance comparison of different network structures. f_m, r_s, r_t denote features of segmentation mask, spatial reference and temporal reference respectively; $\oplus, \otimes$ denote the feature concatenation and correlation operations respectively. Networks are trained only with affine transformation for saving training time.

	MAE	RMSE	PLCC	SRCC
$net1 - f_m \oplus r_s$	0.067	0.103	0.920	0.871
$net2 - f_m \otimes r_s$	0.055	0.08	0.941	0.907
$net3 - f_m \oplus r_t$	0.106	0.155	0.805	0.761
$net4 - f_m \otimes r_t$	0.210	0.261	−0.003	−0.03
$net5 - f_m \otimes (r_s \oplus r_t)$	0.188	0.253	−0.015	−0.01
$net6 - (f_m \otimes r_s) \oplus (f_m \otimes r_t)$	0.125	0.164	0.777	0.756
$net7 - f_m \oplus r_s \oplus r_t$	0.055	0.088	0.941	0.907
$net8 - f_m \oplus r_t \oplus (f_m \otimes r_s)$	0.039	0.061	0.972	0.950
$net9 - r_t \oplus (f_m \otimes r_s)$ *VOSE-Net*	**0.039**	**0.06**	**0.973**	**0.951**

From Table 4, we can see that for single reference networks ($net1 - net4$), r_s (the original frame image) has much better reference value than r_t (motion information): the MAE of net_3 is 58% higher(worse) than that of net_1. We can also see that it is the close correlation between segmentation mask the original image that helps estimate the mask quality: the correlation operation between f_m and r_s helps improve the prediction accuracy (*net*1 vs. *net*2), while the same operation between f_m and r_t heavily harms the overall performance (*net*3 vs. *net*4 or *net*2 vs. *net*4). This phenomenon is further verified by networks with two references which also involve $f_m \otimes r_t$, i.e. $net5, net6$.

Double-reference networks show consistent improvement over single-reference ones, e.g *net*7 decreases 18% in MAE and 15% in RMSE than *net*1 with adding in the temporal space reference; *net*7 decreases 48% in MAE and 43% in RMSE than *net*3 with adding in the spatial space reference. Nevertheless, adding in more features for fusion does not promise better performance, e.g. *net*8 has one more f_m concatenated with $r_t \oplus (f_m \otimes r_s)$ than *net*9, but has lower performance.

The proposed *VOSE-Net* (*net*9) has the best performance of 0.039 MAE, 0.06 RMSE, 0.973 PLCC and 0.951 SRCC, showing that it has a favorable reference type and feature fusion strategy for the video segmentation mask quality evaluation task. Fig. 6 presents some visual examples of segmentation proposals (not in the VISA training or testing set) with the two scores (ground-truth s_j and predicted s_p). The prediction score s_p is very close to s_j, validating the automatic quality evaluation ability of the *VOSE-Net*.

General applicability

To validate the general applicability of the *VOSE-Net*, we directly apply it on various datasets and compare its performance scores. We

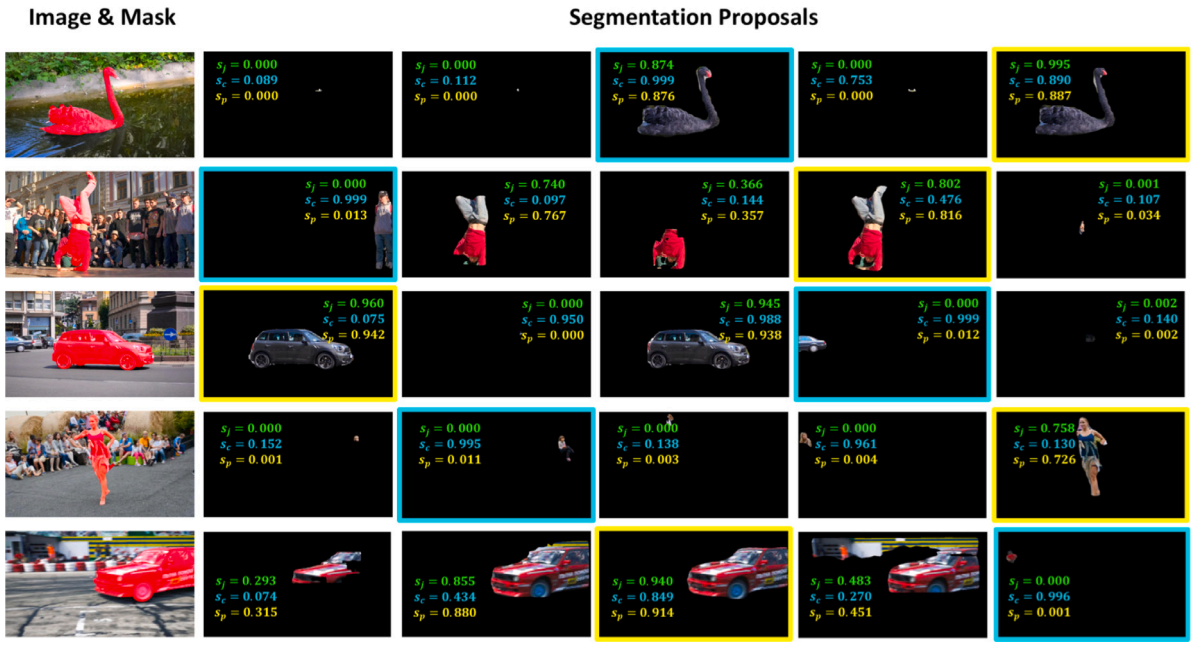

Fig. 6. Example results of the *VOSE-Net* prediction scores on various segmentation proposals. s_j, s_c and s_p denote the ground-truth score, proposal confidence score, and the *VOSE-Net* prediction score, respectively. Proposals with the highest s_p are highlighted in yellow; while proposals with the highest s_c are highlighted in blue.

firstly separate the 'VISA' test set into 'VISA-1' and 'VISA-2' by whether the masks of the algorithms are used in constructing 'VISA' training set or not. As shown in Table 5, performances of the *VOSE-Net* is only slightly different among these three settings, e.g. MAEs for 'VISA', 'VISA-1','VISA-2' are 0034, 0.033 and 0.036, respectively. This demonstrates that the *VOSE-Net* is not overfitted to masks produced by specific algorithms.

We also apply the *VOSE-Net* on the precomputed segmentation results (including intermediate ones) of three new VOS algorithms which are not incorporated in dataset construction, i.e. STCN [23], HMMN [22] and MiVOS [21]. Dataset 'DAVIS' in Table 5 denotes the DAVIS validation set with segmentation masks from the above three algorithms respectively. Table 5 illustrates that the *VOSE-Net* has similar performances on the 'DAVIS' dataset setting with 'VISA', with lower MAE on 'VISA' (0.041 vs. 0.034) and lower RMSE (0.034 vs. 0.052) on 'DAVIS', this may be because that original videos in the VISA dataset are the same as those in the DAVIS database.

Therefore, we further apply the *VOSE-Net* on the much larger YoutubVOS dataset [67]. We randomly select 500 videos with primary instances from the YoutubVOS training set and test the effectiveness of the *VOSE-Net* on large-scale unseen video data. Optical flow for the temporal reference is computed with PWC-Net [44] at both 5 fps and 30 fps, and candidate segmentation masks are generated with [4,68] at both frame rates. As shown in the last two rows of Table 5, the *VOSE-Net* maintains the ability to blindly estimate mask qualities on Youtube videos (with MAE of 0.091). In addition, we observe that the mask quality predictions with 30 fps flows as temporal references are better than those with 5 fps flows (MAE 0.091 vs. 0.112), this shows that the fineness of optical flows can influence the mask quality estimation results (i.e. flows computed at 5 fps are coarser and have more flaws). Another interesting finding is that the most influenced evaluation metric during dataset changing is the SRCC, which measures the correlation between variable data rankings, this shows that when applied to videos different from training data, the *VOSE-Net* may score low-quality masks with high scores and disrupt the score rankings in some cases. We show some visual examples of such error cases in the supplementary material for better illustration.

Table 5
Quality estimation of the *VOSE-Net* on various datasets.

Dataset	MAE ↓	RMSE ↓	PLCC ↑	SRCC ↑
VISA	0.034	0.052	0.980	0.958
VISA-1	0.033	0.051	0.980	0.959
VISA-2	0.036	0.055	0.979	0.956
DAVIS	0.041	0.034	0.944	0.804
Youtube-5fps	0.112	0.125	0.880	0.724
Youtube-30fps	0.091	0.144	0.910	0.715

5.2. Applications

To demonstrate the robustness and practical applicability of the proposed *VOSE-Net*, we apply it to several useful tasks, i.e. segmentation proposal selection, parameter optimization and raw video segmentation evaluation. The first two tasks are carried out on the DAVIS 2016 validation set, where we can use the ground-truth labels to check the effectiveness of the *VOSE-Net*. As for the third task, we download ten web videos which have predominant instances, get segmentation masks from three different video segmentation algorithms, and let the *VOSE-Net* judge the quality of each prediction mask to see if it matches human intuitions. Detailed results and analysis are as follows.

Segmentation proposal selection

One simple application of our quality evaluator is to automatically select the best segmentation proposal from groups of candidates. For example, we extract a bunch of segmentation masks from MaskRCNN [69] with various backbone networks (ResNet50 [70], ResNet101 [70], ResNext101 [71], ResNext152 [71]) and training strategies (1x, 2x, 3x, 4x). For each backbone and training setting (e.g. R50-1x, ResNet50 with strategy 1x), MaskRCNN generates several segmentation masks for each frame with detection confidence scores from 0 to 1. Taking the mask with the highest confidence score as its segmentation output for each image, the performance of each network is as shown in Table 6, row 1–9. When combining all the proposals together, the highest score selection strategy (max_{conf}) does not improve over a single network (e.g. the max_{conf} has J mean of 0.575, which is lower than its subset $X101{-}2x$). Neither does the max_{area} strategy which picks out the largest mask proposal for each frame. In contrast, the *VOSE-Net* can accurately

Fig. 7. Example of the *VOSE-Net* quality estimation for masks on arbitrary videos. This figure presents samples of Internet video frames with segmentation masks from different methods, and the *VOSE-Net* is used to assess mask qualities with its prediction score s_p. High-quality and low-quality masks selected by s_p are shown on the left and right respectively.

assess the quality of each proposal and select the ones most likely to be target masks. From the last row of Table 6, we can see that with the same set of proposals, the *VOSE-Net* proposal selection strategy overwhelms the network confidence strategy (max_{conf}) by 22.3% J mean and 21% F mean. For better visual illustrations, we also show some typical proposals in Fig. 6 with their ground-truth scores s_j (Jaccard similarity computed with ground-truth mask), network confidence scores s_c and the *VOSE-Net* prediction scores s_p, demonstrating that the *VOSE-Net* automatic assessment of each proposal is similar to ground-truth scores without the need for manual labels.

Post-processing parameter optimization

Post-processing techniques are widely used in segmentation tasks [16,72,73] to enhance mask details. For example, [4] adds CRF optimization after video segmentation; [72] employs denseCRF which can be embedded in convolutional network; and [5] trains one mask refinement network for each target as post-processing. However, most post-processing methods need to fine-tune on the first frame mask, or need to grid search for optimized parameter settings with access to the first frame annotation. For unsupervised testing where the videos have no manual labels, empirical values are directly applied to every video despite of their differences. In this case, the optimization process largely relies on researchers' experience. However, we propose that with help of the *VOSE-Net*, this parameter selection process can be done automatically with a thorough exploration of optimization potentials.

Take CRF for an example, there are usually four tunable parameters in the CRF optimization process, the color-dependent terms θ_α and θ_β, the color-independent term θ_γ, and the optimization iteration number $\#Iters$. The commonly suggested values for these four parameters are 80, 13, 3 and 5. We use the proposals selected by the *VOSE-Net* in Table 6 as a baseline algorithm, and conduct a parameter search.

As shown in Table 7, not all the parameter settings are suitable for video segmentation mask post-processing. For $GridSearch$, we search all the parameter settings on the first frame of each test video, and select the one with the highest J mean. For *VOSE-Net*, we search all the parameters on three random frames from each video, and select the setting with the highest prediction score. The last two rows of Table 7 show that parameter selection with the *VOSE-Net* outperforms all candidate parameter settings, as well as the $GridSearch$ strategy which uses first frame annotations. This verifies that the proposed *VOSE-Net* is qualified for an alternative of manual parameter tuning experience.

Table 6
Segmentation proposal selection on the DAVIS 2016 val set.

Method	J mean ↑	J recall ↑	F mean ↑	F recall ↑
R50-1x	0.518	0.592	0.519	0.561
R50-2x	0.550	0.614	0.559	0.585
R101-1x	0.548	0.629	0.555	0.588
R101-2x	0.560	0.637	0.568	0.597
X101-1x	0.565	0.646	0.571	0.608
X101-2x	0.584	0.661	0.594	0.633
X101-3x	0.509	0.572	0.515	0.534
X101-4x	0.576	0.648	0.588	0.606
X152	0.550	0.625	0.557	0.585
max_{area}	0.479	0.563	0.460	0.465
max_{conf}	0.575	0.646	0.586	0.519
VOSE-Net	0.798	0.955	0.791	0.875

Table 7
CRF parameter optimization on the DAVIS 2016 val set.

$[\theta_\alpha, \theta_\beta, \theta_\gamma]$	$\#Iters$ (J mean, F mean)			
	3	5	10	15
[20, 13, 3]	0.811,0.794	0.812,0.795	0.812,0.794	0.812,0.794
[40, 13, 3]	0.806,0.795	0.801,0.788	0.794,0.781	0.790,0.778
[60, 13, 3]	0.785,0.771	0.772,0.757	0.757,0.741	0.751,0.735
[80, 13, 3]	0.763,0.739	0.743,0.719	0.719,0.695	0.708,0.685
[100, 13, 3]	0.712,0.689	0.689,0.659	0.652,0.628	0.637,0.618
[20, 13, 5]	0.748,0.729	0.728,0.706	0.703,0.684	0.691,0.673
[20, 13, 7]	0.741,0.722	0.720,0.699	0.695,0.676	0.684,0.666
[20, 13, 9]	0.735,0.715	0.713,0.692	0.688,0.668	0.677,0.659
[20, 7, 3]	0.780,0.757	0.766,0.746	0.752,0.734	0.745,0.729
[20, 9, 3]	0.771,0.751	0.755,0.736	0.738,0.720	0.729,0.714
[20, 13, 3]	0.811,0.794	0.812,0.795	0.812,0.794	0.812,0.794
[20, 15, 3]	0.747,0.726	0.725,0.701	0.699,0.676	0.687,0.664
Baseline	J mean = 0.798, F mean = 0.791			
$GridSearch$	J mean = 0.801, F mean = 0.788			
VOSE-Net	J mean = 0.812, F mean = 0.803			

Quality evaluation on arbitrary videos

To demonstrate the robustness and universality of the *VOSE-Net*, we download ten videos from the Internet, including real-world, 3D animation and 2D animation movie clips with primary objects of persons, robots, animals, vehicles, etc. We use PWC-Net [44] to compute optical flow for each video, and three different types of segmentation methods to generate video masks [4,69,74]. [69] is a detection-based segmentation network that generates segmentation masks along with detection boxes, where we use the same setting as X101-2x in Table 6 (select the proposals with max confidence scores); [74] is a pixel difference based foreground segmentation method, which computes a dynamic threshold to distinguish foreground pixels from background pixels; and [4] is a fully convolutional neural network which follows the general segmentation idea of FCN [75] and integrates optical flow information within the network.

Fig. 7 shows some segmentations with high qualities (left) and low qualities (right) judged by the *VOSE-Net*. Although we cannot obtain the exact ground-truth scores for these video frames, we can see that the segmentations are in line with human intuitions. For more performance of the *VOSE-Net* on raw videos, we provide a supplementary video for evaluation of the above three methods on all ten videos.

6. Conclusion

In this paper, we build an automatic quality measurement algorithm for video object segmentation masks, the *VOSE-Net*, to accurately estimate mask qualities without access to manual labels. We construct a VIdeo Segmentation Assessment (VISA) dataset to train the proposed evaluator, as well as validate its high precision of quality evaluation. Besides, we demonstrate the robustness of the *VOSE-Net* in three different applications, illustrating its general applicability and universality. We propose that this automatic quality evaluator can be applied widely in field tests and other video segmentation related techniques, helping to assess mask qualities and tune dynamic parameters online without manual intervention.

CRediT authorship contribution statement

Jingchun Cheng: Conceptualization, Methodology, Software, Writing – original draft. **Jiajie Song:** Investigation, Data curation. **Rui Xiong:** Validation, Writing – review & editing. **Xiong Pan:** Supervision, Project administration. **Chunxi Zhang:** Supervision, Funding acquisition.

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 is supported by the China Postdoctoral Science Foundation under Grant No. 2021M690293.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.measurement.2022.111003.

References

[1] Yanchao Yang, Brian Lai, Stefano Soatto, Dystab: Unsupervised object segmentation via dynamic-static bootstrapping*, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2825–2835.
[2] Sabarinath Mahadevan, Ali Athar, Aljosa Osep, Sebastian Hennen, Laura Leal-Taixé, B. Leibe, Making a case for 3D convolutions for object segmentation in videos, ArXiv (2020) abs/2008.11516.
[3] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Computer Vision and Pattern Recognition, 2016.
[4] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, Ming-Hsuan Yang, Segflow: Joint learning for video object segmentation and optical flow, in: ICCV, 2017.
[5] Jonathon Luiten, Paul Voigtlaender, Bastian Leibe, Premvos: Proposal-generation, refinement and merging for video object segmentation, in: ACCV, 2018.
[6] Yeong Jun Koh, Chang-Su Kim, Primary object segmentation in videos based on region augmentation and reduction, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 7417–7425.
[7] Suyog Dutt Jain, Bo Xiong, Kristen Grauman, Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2117–2126.
[8] Won-Dong Jang, Chang-Su Kim, Online video object segmentation via convolutional trident network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5849–5858.
[9] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, Liang-Chieh Chen, Feelvos: Fast end-to-end embedding learning for video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9481–9490.
[10] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, Ming-Hsuan Yang, Fast and accurate online video object segmentation via tracking parts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7415–7424.
[11] Paul Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.
[12] Fanman Meng, Lili Guo, Qingbo Wu, Hongliang Li, A new deep segmentation quality assessment network for refining bounding box based segmentation, IEEE Access 7 (2019) 59514–59523.
[13] W. Shi, Fanman Meng, Q. Wu, Segmentation quality evaluation based on multi-scale convolutional neural networks, 2017 IEEE Visual Communications and Image Processing (VCIP) (2017) 1–4.
[14] C. Huang, Q. Wu, Fanman Meng, Qualitynet: Segmentation quality evaluation with deep convolutional networks, 2016 Visual Communications and Image Processing (VCIP) (2016) 1–4.
[15] J. Pont-Tuset, F. Marqués, Supervised evaluation of image segmentation and object proposal techniques, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 1465–1478.
[16] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, Luc Van Gool, One-shot video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 221–230.
[17] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, Michael Felsberg, A generative appearance model for end-to-end video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8953–8962.
[18] Nicolas Märki, Federico Perazzi, Oliver Wang, Alexander Sorkine-Hornung, Bilateral space video segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 743–751.
[19] Matthias Grundmann, Vivek Kwatra, Mei Han, Irfan Essa, Efficient hierarchical graph-based video segmentation, in: 2010 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2141–2148.
[20] Federico Perazzi, Oliver Wang, Markus Gross, Alexander Sorkine-Hornung, Fully connected object proposals for video segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3227–3234.
[21] Ho Kei Cheng, Yu-Wing Tai, Chi-Keung Tang, Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5555–5564.
[22] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, Euntai Kim, Hierarchical memory matching network for video object segmentation, in: ICCV, 2021.
[23] Ho Kei Cheng, Yu-Wing Tai, Chi-Keung Tang, Rethinking space-time networks with improved memory coverage for efficient video object segmentation, in: NeurIPS, 2021.
[24] Paul Voigtlaender, Bastian Leibe, Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation, in: The 2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops, Vol. 5, 2017.
[25] Zongxin Yang, Yunchao Wei, Yi Yang, Collaborative video object segmentation by multi-scale foreground-background integration, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1.
[26] Xiaoxiao Li, Yuankai Qi, Zhe Wang, K. Chen, Z. Liu, J. Shi, Ping Luo, X. Tang, Chen Change Loy, Video object segmentation with re-identification, ArXiv (2017) abs/1708.00197.

[27] Xiaoliang Hu, Zhijiang Xie, Fei Liu, Assessment of speckle pattern quality in digital image correlation from the perspective of mean bias error, Measurement (2020) 108618.
[28] Michal Kedzierski, Damian Wierzbicki, Radiometric quality assessment of images acquired by uav's in various lighting and weather conditions, Measurement 76 (2015) 156–169.
[29] Deepti Ghadiyaram, A. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Trans. Image Process. 25 (2016) 372–387.
[30] Patrick Le Callet, Florent Autrusseau, Subjective quality assessment IRCCyN/IVC database, 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.
[31] V. Movahedi, J. Elder, Design and perceptual validation of performance measures for salient object segmentation, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (2010) 49–56.
[32] G. Csurka, Diane Larlus, F. Perronnin, What is a good evaluation measure for semantic segmentation? in: BMVC, 2013.
[33] M. Everingham, S. Eslami, L. Gool, C.K. Williams, J. Winn, Andrew Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vis. 111 (2014) 98–136.
[34] K. Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, CoRR abs/1409.1556.
[35] Zifeng Wu, Chunhua Shen, Anton Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, Pattern Recognit. 90 (2019) 119–133.
[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., Imagenet large scale visual recognition challenge, IJCV (2015).
[37] Zhuxin Chen, Zhifeng Xie, Weibin Zhang, Xiangmin Xu, Resnet and model fusion for automatic spoofing detection, in: INTERSPEECH, 2017, pp. 102–106.
[38] Jifeng Dai, Yi Li, Kaiming He, Jian Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems, 2016, pp. 379–387.
[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[40] Matthew D Zeiler, Rob Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[41] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, Thomas Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[42] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, Thomas Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: CVPR, 2017.
[43] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, Cordelia Schmid, Epicflow: Edge-preserving interpolation of correspondences for optical flow, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1164–1172.
[44] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, Jan Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: CVPR, 2018.
[45] Tim Meinhardt, Laura Leal-Taixé, Make one-shot video object segmentation efficient again, in: NeurIPS, 2020.
[46] Hongje Seong, Junhyuk Hyun, Euntai Kim, Kernelized memory network for video object segmentation, in: ECCV, 2020.
[47] Qingnan Fan, Fan Zhong, Dani Lischinski, Daniel Cohen-Or, Baoquan Chen, Jumpcut: non-successive mask transfer and interpolation for video cutout, ACM Trans. Graph. 34 (6) (2015) 195:1–195:10.
[48] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, Bernt Schiele, Lucid data dreaming for video object segmentation, Int. J. Comput. Vis. (2018) 1–23.
[49] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, Alexander Sorkine-Hornung, Learning video object segmentation from static images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2663–2672.
[50] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, In So Kweon, Pixel-level matching for video object segmentation using convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2167–2176.
[51] S Avinash Ramakanth, R Venkatesh Babu, Seamseg: Video object segmentation using patch seams, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 376–383.
[52] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, Philip HS Torr, Fast online object tracking and segmentation: A unifying approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1328–1338.
[53] Anna Khoreva, Anna Rohrbach, Bernt Schiele, Video object segmentation with language referring expressions, in: Asian Conference on Computer Vision, Springer, 2018, pp. 123–141.
[54] Varun Jampani, Raghudeep Gadde, Peter V Gehler, Video propagation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 451–461.
[55] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, Haibin Ling, Learning unsupervised video object segmentation through visual attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3064–3074.
[56] Margret Keuper, Bjoern Andres, Thomas Brox, Motion trajectory segmentation via minimum cost multicuts, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3271–3279.
[57] Brian Taylor, Vasiliy Karasev, Stefano Soatto, Causal video object segmentation from persistence of occlusions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4268–4276.
[58] Anestis Papazoglou, Vittorio Ferrari, Fast object segmentation in unconstrained video, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1777–1784.
[59] Yong Jae Lee, Jaechul Kim, Kristen Grauman, Key-segments for video object segmentation, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1995–2002.
[60] Pavel Tokmakov, Karteek Alahari, Cordelia Schmid, Learning video object segmentation with visual memory, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4481–4490.
[61] Pavel Tokmakov, Cordelia Schmid, Karteek Alahari, Learning to segment moving objects, Int. J. Comput. Vis. 127 (3) (2019) 282–301.
[62] Mennatullah Siam, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, Martin Jagersand, Video segmentation using teacher-student adaptation in a human robot interaction (HRI) setting, 2018, arXiv preprint arXiv:1810.07733.
[63] Peter Ochs, Thomas Brox, Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1583–1590.
[64] Brent Griffin, Jason Corso, Tukey-inspired video object segmentation, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1723–1733.
[65] Katerina Fragkiadaki, Geng Zhang, Jianbo Shi, Video segmentation by tracing discontinuities in a trajectory embedding, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1846–1853.
[66] Yi Yang, Yueting Zhuang, Yunhe Pan, Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies, Front. Inf. Technol. Electron. Eng. (2021).
[67] N. Xu, L. Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott D. Cohen, Thomas S. Huang, Youtube-VOS: Sequence-to-sequence video object segmentation, in: ECCV, 2018.
[68] S. Oh, Joon-Young Lee, N. Xu, S. Kim, Video object segmentation using space-time memory networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9225–9234.
[69] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross B. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 386–397.
[70] Kaiming He, X. Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[71] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, Aggregated residual transformations for deep neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5987–5995.
[72] Philipp Krähenbühl, V. Koltun, Efficient inference in fully connected CRFs with Gaussian edge potentials, in: NIPS, 2011.
[73] K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, Video object segmentation without temporal information, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 1515–1530.
[74] Leyi Xiao, H. Ouyang, Chaodong Fan, An improved otsu method for threshold segmentation based on set mapping and trapezoid region intercept histogram, Optik 196 (2019) 163106.
[75] J. Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

Jingchun Cheng is a Post-Doc at BeiHang University. She received her bachelor's degree in Electronic Engineering at Tsinghua University in 2014, was a visiting scholar in UC Merced in 2017, and received her Ph.D. degree in Electronic Engineering from Tsinghua University in 2020. Her research interests include video object segmentation, face recognition, autonomous driving, etc.