SPFTN: A Self-Paced Fine-Tuning Network for Segmenting Objects in Weakly Labelled Videos

Dingwen Zhang[1], Le Yang[1], Deyu Meng[2], Dong Xu[3], Junwei Han[1]*
[1]Northwestern Polytechincal University, [2]Xi'an Jiaotong University, [3]University of Sydney
zdw2006yyy@mail.nwpu.edu.cn, nwpuyangle@gmail.com, dymeng@mail.xjtu.edu.cn
dong.xu@sydney.edu.au, junweihan2010@gmail.com

Abstract

Object segmentation in weakly labelled videos is an interesting yet challenging task, which aims at learning to perform category-specific video object segmentation by only using video-level tags. Existing works in this research area might still have some limitations, e.g., lack of effective DNN-based learning frameworks, under-exploring the context information, and requiring to leverage the unstable negative video collection, which prevent them from obtaining more promising performance. To this end, we propose a novel self-paced fine-tuning network (SPFTN)-based framework, which could learn to explore the context information within the video frames and capture adequate object semantics without using the negative videos. To perform weakly supervised learning based on the deep neural network, we make the earliest effort to integrate the self-paced learning regime and the deep neural network into a unified and compatible framework, leading to the self-paced fine-tuning network. Comprehensive experiments on the large-scale YouTube-Objects and DAVIS datasets demonstrate that the proposed approach achieves superior performance as compared with other state-of-the-art methods as well as the baseline networks and models.

1. Introduction

With the rapidly growing popularity of the video sharing social media (e.g., YouTube), a massive amount of videos can be easily accessed online. This offers the vision community an exciting opportunity to learn visual concepts and object models from the real-world online videos [27]. However, it is hard to directly exploit these online videos in traditional ways because most of online videos are weakly labelled [30, 10]. These videos are only associated with semantic tags to indicate the main objects or concepts within them, whereas the detailed spatial-temporal segmentation masks are not provided due to the heavy burden of manual annotation. Thus, in this paper, we focus on the task of segmenting objects in weakly labelled videos. This task is of great significance in two-fold reasons. On one hand, it could help automatically provide the spatial-temporal segmentation annotations for the online videos so that these online resources can be utilized to help other tasks like training classifiers for image classification [24, 26]. On the other hand, it could act as an essential step towards video content understanding and thus improve the performances of other tasks like video summarization [33] and event detection [3].

In order to segment objects in weakly labelled videos, the most pioneering attempt by Hartmann et al. [8] formulated it as learning weakly supervised classifiers for a set of independent spatial-temporal segments and utilized the graph cuts to refine the obtained object seeds to generate the final object masks. Afterwards, Tang et al. [27] presented a Concept Ranking According to Negative Exemplars (CRANE) algorithm, which is robust to label noise and highly parallelizable and thus could effectively handle large amounts of video data and spatial-temporal segments. Liu et al. [18] presented a nearest neighbor-based label transfer scheme for weakly supervised video segmentation, which mainly focused on the challenging multi-class video segmentation problem. More recently, Zhang et al. [39] proposed a segmentation-by-detection framework, where object and region detectors pre-trained on still images were used to generate the detection and segmentation proposals. Then, object trackers were refined by inferring shape likelihoods to suppress background noise while preserving the spatial-temporal consistency of foreground objects.

As can be seen, the existing works usually first decompose the positive and negative videos into a number of spatial-temporal segments. Then the segmentation-level classifiers or inference models are trained under the weak supervision to identify the segments related to the given object categories in videos. Finally, post-processing methods are applied to refine the object segmentation masks. Although approaches along this pipeline have achieved good

*Corresponding author.

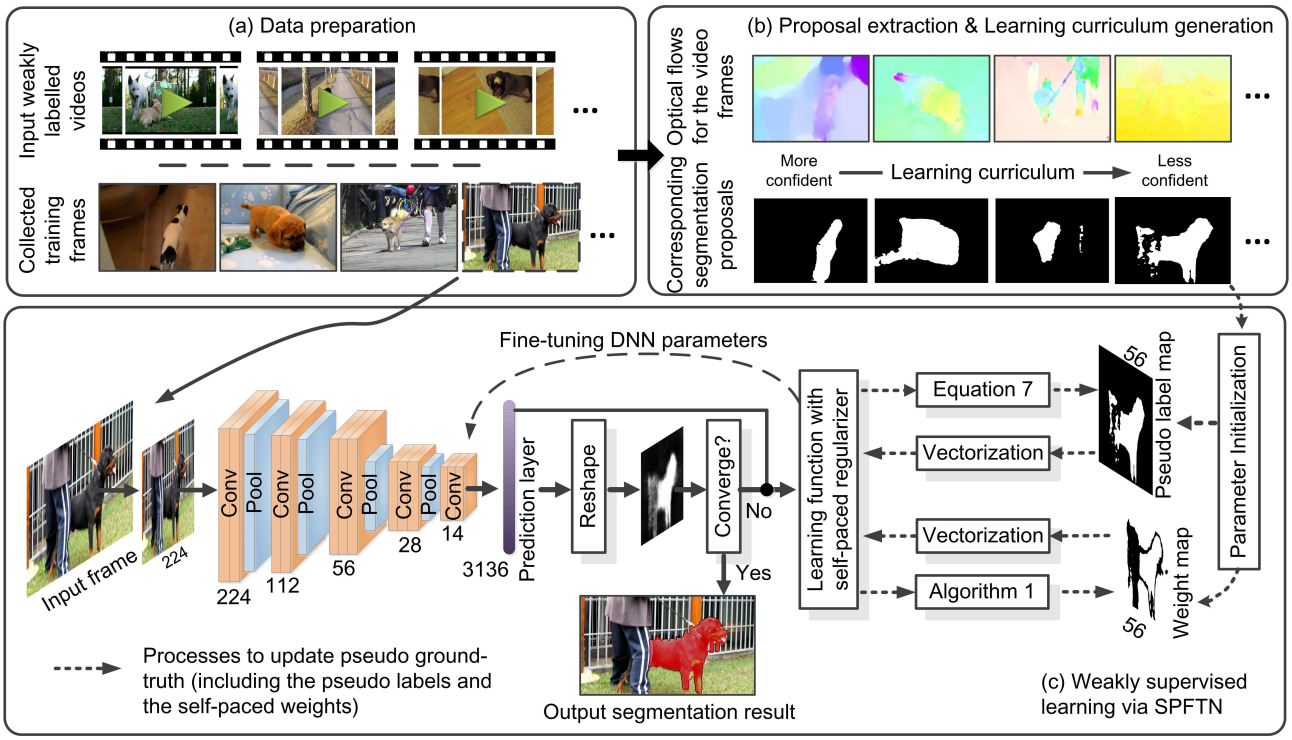

Figure 1. The proposed self-paced fine-tuning network-based framework for object segmentation in weakly labelled videos. Equipped with the newly proposed self-paced regularizer, the network can not only generate the pseudo label map to provide the pixel-level pseudo labels but also the weight map to indicate the reliable pixels during the learning process, which could work effectively under weak supervision.

performance in various cases, there might still be some limitations that could be addressed for further improvement. First, it is unclear how to address the investigated problem via DNN-based frameworks, which have shown excellent performance in many other computer vision problems. Second, most of the existing methods consider each spatial-temporal segment as an individual instance during their learning processes. Thus, the scene context in each frame, which can provide helpful contextual priors for object recognition [29], remains under-exploited in this area. Third, most existing methods require not only the positive videos but also the negative videos. However, although the negative videos can be easily collected, principle ways to determine the quantity and quality of them are not well studied, leading to unstable performance of the final results.

In order to tackle the aforementioned limitations, we propose a novel self-paced fine-tuning network (SPFTN) in this paper. As shown in Fig. 1, given a group of videos that are weakly labelled as containing common objects from one semantic category, the proposed approach first prepares training data by decomposing these videos into frames and generating segmentation proposals for these frames. Then, a unified learning process is proposed to segment semantic objects within the videos. In the proposed network, we utilize a fully connected layer before final prediction, which guarantees that the receptive field of each output node is the entire input video frame. Thus, the labels of each pixel could be inferred with perception of the global structure of each input frame scene, which encodes rich context information. In addition, by learning the object segmentation masks in this way, we could obtain satisfactory performance by only using a collection of positive videos. Thus the proposed approach requires less manual efforts to collect the negative videos, which also resolves the instability issue caused by the negative videos.

Essentially, one of the most critical issue in the proposed framework is how to design effective deep neural network (DNN) under weak supervision. As we know, DNN has achieved tremendous success in various problems like object detection [6] and saliency prediction [7]. However, in most cases, the DNN needs to be trained under fully supervision, while training DNN under weakly supervision remains to be challenging and under-addressed, especially for the video object segmentation task. To address this problem, we propose to incorporate the self-paced learning regime into the DNN fine-tuning process to cope with the data ambiguity problem and guide an effective learning manner in complex scenarios. It thus leads to the novel SPFTN. Specifically, inspired by the learning process of humans/animals, the theory of self-paced (or curriculum) learning [1, 15] is proposed in recent years. The main idea is to learn the model iteratively from easy to complex

samples in a self-paced fashion. The effectiveness of such learning regime, especially its effectiveness in highly ambiguous data, has been validated in various computer vision tasks [36, 13, 40]. Among the existing works, all the established self-paced learning regimes were designed based on the conventional shallow learning models like support vector machine (SVM), whereas little successful attempt has been made to integrate such effective learning regime with more powerful deep models. Consequently, it motivates us to make the earliest effort to design proper mechanism to integrate the self-paced learning and DNN into a unified and compatible framework. On one hand, our work can further improve the learning capability of self-paced regime and. On the other hand, our work also performs weakly supervised training of DNN parameters. In addition, for improving the effectiveness of the self-paced learning regime, we propose to introduce a novel group curriculum term into the optimization objective, which could leverage helpful prior-knowledge to guide the learner to select confident training samples while considering the group-level learning priority flexibly. Compared with that in instance-level, the learning priority in group-level tends to be much cheaper yet more effective to guide the learning procedure.

To sum up, this paper has three main contributions:

1) This paper proposes a novel SPFTN approach for weakly supervised video object segmentation, which is carefully designed to integrate the self-paced learning regime and the DNN learning function into a unified and compatible framework. It could improve the learning capability of the self-paced regime and perform weakly supervised training of DNN model.

2) To better leverage the helpful prior-knowledge, we propose a novel self-paced regularizer by introducing the group curriculum term into the optimization problem. The group curriculum term imposes two principles for selecting confident training samples, which could enable the learner to consider the sample priority and diversity flexibly.

3) The proposed learning framework can effectively encode rich context information during the learning process and capture adequate object semantics only from the positive videos, which helps improve the segmentation accuracy and increase learning stability, respectively.

2. Self-Paced Fine-Tuning Network

2.1. Network Architecture

The network is established based on the VGG 16 network [25] with modified objective function and additional pseudo label layer as well as weight layer for implementing the self-paced fine-tuning under weak supervision. As shown in Fig. 1, it consists of 20 layers, including 13 convolutional layers (the orange layers), 4 max-pooling layers (the blue layers), 1 fully connected prediction layer (the purple layer), 1 pseudo label layer, and 1 weight layer. We use raw video frames resized to 224×224 pixels as the network inputs. The network first adopts the first 13 convolutional layers and 4 max-pooling layers as in the VGG 16 network [25] to extract deep features of each frame. Then a fully connected layer with 3136 nodes is used to predict the segmentation map of size 56×56. We also introduce a pseudo label layer and a weight layer with 3136 nodes (obtained by vectorizing the 56×56 demential pseudo label map and weight map, respectively) to provide pseudo supervision to guide the fine-tuning of the entire network. The convolutional layers can gradually involve relatively larger receptive field of context information into learning and the fully connected layer helps to encode the global context into the final prediction. Thus, it is able to leverage rich context information for inferring the object segmentations.

2.2. Objective Function

Given a collection of K video frames $\{I_k\}_{k=1}^{K}$ extracted from a set of weakly labelled videos from one semantic category, the input dimension of the designed network architecture is set to be 244×244. Corresponding to each input frame, the pseudo labels are denoted as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K] \in \{-1, 1\}^{d \times K}$, where $\mathbf{y}_k \in \{-1, 1\}^d$ denotes the structural pseudo label of I_k (background pixels are labeled as 0 and vice versa) and $d = 3136$ is the output dimension of the network. As the pseudo labels can be easily transformed to the pseudo ground-truth masks by reshaping, they can provide supervision to the designed network. In order to enable the network to effectively work under weak supervision, we introduce the self-paced learning regime into parameter fine-tuning. With the input of the video frames and the initial $\mathbf{Y}$ and $\mathbf{V}$, the learning objective gradually discovers confident training samples and use them to fine-tune DNN via mainly minimizing a weighted prediction loss term and a self-paced regularizer:

$$
\begin{aligned}
&\min_{\mathbf{W},\mathbf{Y},\mathbf{V}} \ \mathbf{E}(\mathbf{W}, \mathbf{Y}, \mathbf{V}) = \\
&\quad r(\mathbf{W}) + \sum_{k=1}^{K} L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W})) + f(\mathbf{V}; \mathbf{p}, \lambda, \gamma, \tau), \\
&s.t. \ \ \mathbf{V} \in [0,1]^{d \times K}, \mathbf{p} \in [0,1]^K \\
&\quad \textstyle\sum_k ||\mathbf{v}_k||_1 \in (0, d \times K), \\
&\quad \textstyle\sum_k ||\mathbf{y}_k||_1 \in (0, d \times K).
\end{aligned}
\tag{1}
$$

Here $r(\cdot)$ indicates the squared ℓ_2 norm, $\mathbf{W}$ indicates the trainable parameters among the network, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K]$ denotes the weight matrix which reflects the self-paced weights for all the pixels of the video frames, $\mathbf{v}_k \in [0,1]^{d \times 1}$, λ, γ, and τ are parameters for controlling the learning pace, $\mathbf{p} = [p_1, p_2, \ldots, p_K]$ is the curriculum

variable encoding the learning priority of each video frame, and $\Phi(I_k|\mathbf{W})$ indicates the prediction function of the network, which forward propagates I_k to the prediction layer via the network parameters $\mathbf{W}$. $L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W}))$ indicates the weighted hinge loss:

$$L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W})) = \sum_{i=1}^{d} v_k^i \max(1 - y_k^i \cdot \Phi(I_k|\mathbf{W})^i, 0)^2, \tag{2}$$

where v_k^i, y_k^i, and $\Phi(I_k|\mathbf{W})^i$ indicate the i-th dimension of the weight vector $\mathbf{v}_k$, pseudo label vector $\mathbf{y}_k$, and prediction vector $\Phi(I_k|\mathbf{W})$, respectively. As in [4], we adopt the hinge loss in the square form for ease of optimization. For the constraints, the first one defines the range of the variables; the second one indicates that only a part of samples would be selected during the learning procedure; the third one indicates that the input videos contain both foreground and background regions.

In (1), the self-paced learning capability is guided by a learning curriculum which is pre-defined based on the helpful prior-knowledge. This capability is followed by the involvement of a novel self-paced regularizer that consists of a sample easiness term and a group curriculum term:

$$f(\mathbf{V}; \mathbf{p}, \lambda, \gamma, \tau) = \underbrace{-\lambda \sum_{k=1}^{K} ||\mathbf{v}_k||_1}_{Sample\ \ easiness} \underbrace{-\gamma \sum_{k=1}^{K} (\tau + p_k) \sqrt{\sum_{i=1}^{d} v_k^i}}_{Group\ \ curriculum}. \tag{3}$$

Specifically, the sample easiness term, i.e., the negative ℓ_1-norm term, is inherited from the conventional SPL, which favors selecting easy over complex examples. If we omit the group curriculum term (i.e., let $\gamma = 0$), the regularizer degenerates to the traditional hard SPL function proposed in [15], which outputs either 1 or 0 for the weight v_k^i, by judging whether its loss value is smaller than the pace parameter λ or not. That is, a sample with smaller loss is taken as an "easy" sample and thus should be learned with preference and vice versa.

The group curriculum term, i.e., the negative weighted sparsity term, favors selecting training samples by following the pre-defined learning curriculum, which mainly consists of two principles: 1) samples residing in the training frames with high learning priority are likely to be selected in earlier stage; 2) samples concentrating in a limited number of groups (training frames) are not preferred. These principles could be easily understood by rewriting this term as:

$$-\gamma \left(\sum_{k=1}^{K} p_k \sqrt{\sum_{i=1}^{d} v_k^i} + \tau \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{d} v_k^i} \right). \tag{4}$$

In (4), minimizing the first term tends to assign non-zero values of v_k^i to the samples residing in the groups (training frames) with higher priority p_k, and thus it corresponds to the first principle. For the second term, we can see it as the anti-group-wise sparse representation of $\mathbf{V}$, which has a counter-effect to group-wise sparsity [32]. Different from the $\ell_{2,1}$-norm used in [13], the $\ell_{0.5,1}$-norm used here is convex and can lead to real valued solution. Minimizing this term tends to disperse non-zero elements of v_k^i over more groups and thus it corresponds to the second principle. By using such group curriculum term, we can only provide the group (frame)-level curriculum to guide the learning process, which tends to be much easier than providing the instance (pixel)-level curriculum. Moreover, it is also able to provide informative prior to the instances as the pixels in easy image scenes usually tend to have larger confidence.

By using the proposed regularizer $f(\mathbf{V}, \mathbf{p}; \lambda, \gamma, \tau)$, the learner could infer reliable learning pace based on both the knowledge captured by itself and the prior-knowledge provided by the learning curriculum. Different from most of the previous works, the learner in the proposed learning regime takes account of the learning curriculum but is not dogmatically determined by it. This is important as, on one hand, the learning curriculum is defined without considering the knowledge of the learner. So it might not fit to the learner well. On the other hand, the learning curriculum is defined based on general knowledge, which may not perfectly fit into the specific situation, e.g., the specific object category and frame scene, during the learning procedure.

In addition, Eq. (1) does not enforce that at least one positive instance would be emerged in each positive frame, which is different from some weakly supervised learning formulations [36]. This would be helpful for handling the noisy case when the labelled objects do not appear in some video frames. As we know, even in the positive videos, there are still a large amount of background regions that could provide informative negative samples for the learner. Such negative samples possibly provide quite discriminative knowledge for helping distinguish the semantic category from its surrounding context. Thus, the proposed learning regime can only use the frames from positive videos, which can alleviate the labor to collect negative videos and cope with the underlying instability issue.

2.3. Optimization

The solution of Eq. (1) can be approximately obtained via the alternative search strategy which optimizes the parameters $\mathbf{W}, \mathbf{Y}$ and $\mathbf{V}$ alternatively. More specifically, we first initialize $\mathbf{Y}$ and $\mathbf{V}$. Then, the optimization strategy in each iteration consists of the following steps :

*Optimize **W** with fixed **Y** and **V**:* This step aims to update the DNN parameters under the supervision of the pseudo label layer and weight layer. In this case, Eq. (1) degenerates

to the form of:

$$\min_{\mathbf{W}} \ r(\mathbf{W}) + \sum\nolimits_{k=1}^{K} L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W})). \tag{5}$$

This is essentially the conventional objective function of the DNN with different important weights for the training samples. Thus, the DNN parameters $\mathbf{W}$ could be easily optimized by the widely used back-propagation algorithm.

Optimize $\mathbf{Y}$ with fixed $\mathbf{W}$ and $\mathbf{V}$: The goal of this step is to learn the pseudo labels of the pixels within the training frames from the current DNN model. Eq. (1) in this case could be reformulated as:

$$\min_{\mathbf{Y}} \ \sum\nolimits_{k=1}^{K} L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W})), \tag{6}$$

which leads to the following solution as shown in [19, 12]:

$$y_k^i = \underset{y_k^i \in \{+1,-1\}}{arg\min} \ \max(1 - y_k^i \cdot \Phi(I_k|\mathbf{W})^i, 0)^2. \tag{7}$$

Optimize $\mathbf{V}$ with fixed $\mathbf{Y}$ and $\mathbf{W}$: After updating the pseudo labels, we aim to renew the weights on all pixels to differentiate their importance to the learner. In this case, Eq.. (1) is reformulated as:

$$\begin{aligned} &\min_{\mathbf{V}} \ \sum_{k=1}^{K} L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W})) + f(\mathbf{V}; \mathbf{p}, \lambda, \gamma, \tau) \\ = &\min_{\mathbf{V}} \ \sum_{k=1}^{K}\sum_{i=1}^{d} v_k^i \max(1 - y_k^i \cdot \Phi(I_k|\mathbf{W})^i, 0)^2 \\ - &\ \lambda \sum_{k=1}^{K}\sum_{i=1}^{d} v_k^i - \gamma \sum_{k=1}^{K} (\tau + p_k) \sqrt{\sum_{i=1}^{d} v_k^i}, \end{aligned} \tag{8}$$

which becomes a convex optimization problem. Based on the KKT (Karush Kuhn Tucker) conditions, the global optimum of Eq. (8) can be efficiently calculated via Algorithm 1, where the loss term $\max(1 - y_k^i \cdot \Phi(I_k|\mathbf{W})^i, 0)^2$ is simplified as l_k^i.

As can be seen, Eq. (8) leads to the real-valued solution to the weight layer, i.e., the samples whose losses are smaller than the threshold $\lambda + \gamma(\tau + p_k)/(2\sqrt{i})$ would be taken as the most confident samples, and would be assigned with $v_k^i = 1$ to guide the subsequent fine-tuning of DNN, while the samples whose losses are equal to the threshold would also be selected as the training samples to fine-tune DNN but with less confidence, i.e., $v_k^i \in (0, 1)$. Other samples whose losses are larger than the threshold would be considered as the unconfident training samples and would not be selected ($v_k^i = 0$) for the subsequent fine-tuning procedure. In the proposed self-paced regularizer, the parameter λ controls the learning pace, which corresponds to the age of the model physically. When λ is small, only samples with small losses would be considered as the confident ones. As λ grows, more complex samples with larger losses would be gradually involved into the learning procedure to obtain a more mature model. The parameter γ controls the weight of the learning curriculum. A small γ indicates that the learner relies more on its own learning pace, while a larger γ indicates that the learner also values the helpful prior-knowledge brought by the learning curriculum. The parameter τ controls the weight between the two terms in Eq. (4). A small τ indicates that the learner selects samples mainly according to the learning priority, while a larger τ indicates that the learner tends to select samples from more diverse video frames. With these properties, the proposed self-paced regularizer could provide a theoretically sound way to effectively learn helpful information from the pseudo labels under the weak supervision.

Algorithm 1: Algorithm for optimizing $\mathbf{V}$.

input : K video frames $I_1, \cdots, I_K$ with the corresponding learning priority $p_1, \cdots, p_K$, the DNN model $\mathbf{W}$, parameters λ, γ, and τ;
output: Solution $\mathbf{V}$ in Eq. (8);

for $k = 1$ **to** K **do**
  Sort the pixel instances in I_K in ascending order, i.e., $l_k^1 \leq l_k^2 \leq \cdots \leq l_k^d$; Let $m = 0$;
  for $i = 1$ **to** d **do**
    if $l_k^i < \lambda + \gamma(\tau + p_k)/(2\sqrt{i})$
    then $v_k^i = 1$;
    if $l_k^i \geq \lambda + \gamma(\tau + p_k)/(2\sqrt{i})$
    then count the number m where $l_k^j = l_k^i$
    for $j = i, i+1, \cdots, d$,
    let $v_k^i = \cdots = v_k^{i+m-1}$
      $= ((\gamma(\tau + p_k)/2(l_k^i - \lambda))^2 - (i-1))/m$,
    and $v_k^{i+m} = \cdots = v_k^d = 0$;
    Break;
  end
end
return $\mathbf{V}$.

2.4. Detailed Learning Approach

In this section, we introduce the detailed approach for training the SPFTN and generating the final segmentation masks for the category-specific objects in weakly labelled videos. As shown in Fig. 1 (a), we first collect video frames from a set of weakly labelled videos containing a certain type of semantic object. Then, we extract the optical flow from each video frame to capture the motion information and use it to generate the segmentation proposals via [16][1], which is an unsupervised approach and can only generate

[1]Detailed process to extract segmentation proposals could be found in the supplementary material.

Algorithm 2: The overall approach to apply our SPFT-N for object segmentation in weakly labelled videos.

input : Videos weakly labelled as containing a certain type of object;
output: The semantic object segmentation masks for each video frame;

Collect video frames and the corresponding segmentation proposals with data augmentation;
Pre-train the network;
Obtain learning curriculum by calculating p_k;
Initialize the pseudo labels $\mathbf{Y}$, the self-paced weights $\mathbf{V}$, and assign the parameter values λ, γ, and τ;
while *not converge* **do**
    Fine-tune the DNN parameters $\mathbf{W}$ via Eq. (5);
    Update the pseudo labels $\mathbf{Y}$ via Eq. (6);
    Update the self-paced weights $\mathbf{V}$ via Eq. (8);
    Re-augment the training data and update λ;
end
Use the prediction maps obtained in the last iteration to generate the final segmentation masks;
return the fine-tuned DNN model and the object segmentation masks in the given videos.

coarse estimation as shown in Fig. 1 (b). Then, we augment training data by horizontal-flipping and randomly cropping to cope with the potential over-fitting issue. $\mathbf{Y}$ is initialized by using the segmentation proposal, i.e., the pixels within proposal region are 1 and -1 otherwise. The values in $\mathbf{V}$ are equally initialized as 1.

Before learning the network parameters on the collected training data, we pre-train the network on auxiliary data. Different from [39] which utilizes the part-based detectors trained on PASCAL dataset to assist the learning process, we pre-train our model on the MSRA 10K dataset [5] (containing random objects like "flower" and "traffic sign") under the task of saliency detection [7, 17], which could guide the network to encode general saliency priors from the natural stimulus rather than the specific semantic objects appearing in the given video collections. Just like in [30, 37], it has been a natural trend to introduce or transfer helpful knowledge in weakly supervised tasks.

After pre-training, for each video frame, we calculate p_k as the intersection-over-union (IOU) overlap between the obtained segmentation proposal and the binarized saliency mask, which forms the learning curriculum to guide the subsequent self-paced fine-tuning procedure. Here a larger p_k indicates more consistency between the segmentation proposal and saliency mask. Thus, the content of the corresponding video frame tends to be more confident for the subsequent learning procedure. On the contrary, the video frames with smaller p_k tend to be less confident.

Finally, as shown in Fig. 1 (c), we fine-tune the DNN model to generate the segmentation masks for the specific semantic objects appearing in the given video collection. As introduced in Sec. 2.3, the whole fine-tuning process is performed in a self-paced fashion. In the first iteration, given the initial pseudo label map and weight map, we reshape them to the 3136-dimensional vectors for fine-tuning the parameters $\mathbf{W}$ in the DNN. Then $\mathbf{Y}$ and $\mathbf{V}$ can be optimized subsequently to obtain the updated label map and weight map for guiding the learning in the next iteration. The converge condition is set based on the IOU of the predicted segmentation masks[2] between two neighboring iterations, i.e., if the IOU tends to be smaller than a threshold T, the iteration would be terminated. Notice that after each iteration, we re-augment the training data in order to further alleviate over-fitting during the learning process.

Once reaching the converge condition, we put each video frame into the fine-tuned DNN and up-sample the obtained prediction map to the original size of the input frame. To compensate the resolution degeneration during up-sampling, we follow [39] to adopt the graph-cut method. The overall approach is shown in Algorithm 2.

3. Experimental Results

3.1. Datasets and Implementation Details

We performed experiments on two challenging datasets. The first one is the YouTube-Object dataset [27, 9], which is originally collected in [24]. It consists of objects belonging to 10 semantic categories and totally contains 5507 videos (shots) and 571,089 frames. For providing the pixel-level ground-truth annotation, [9] collected fine-grained pixel-level masks of the foreground object in every 10-th frame for each video, which in total yielded more than 20,000 frames that could be used for quantitative evaluation. The second one is the DAVIS dataset [22], which comprises a total of 50 sequences, 3455 annotated frames, all captured at 24fps and HD 480p spatial resolution, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motionblur and appearance changes. Each video is accompanied by per-pixel, per-frame ground truth segmentation. The standard IOU overlap (calculated by comparing the predict segmentation masks and the corresponding ground-truth masks) is adopted to evaluate the experimental results on these datasets.

We implemented the proposed fine-tuning process using the Caffe library [11]. Within each iteration, the steps for tuning $\mathbf{W}$ and the batchsize for batch processing were adjusted according to the quantity of training exemplar, ensuring every exemplar would be learnt five times. The learning rate in the first iteration was set to 5×10^{-7}, and then

[2]The segmentation masks here only indicate those of the raw video frames without the augmentation data.

Table 1. Results on the YouTube-Object dataset in terms of IOU (higher values indicate better results).

	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	Ave.
Tang et al. [27]	0.178	0.198	0.225	0.383	0.236	0.268	0.237	0.140	0.125	0.404	0.239
Zhang et al. [35]	0.597	0.427	0.276	0.465	0.460	0.414	0.470	0.380	0.061	0.366	0.391
Papazoglou et al. [20]	0.674	0.625	0.378	0.670	0.435	0.327	0.489	0.313	0.331	0.434	0.468
Wang et al. [31]	0.771	0.614	0.365	0.629	0.382	0.437	0.453	0.440	0.243	0.434	0.477
Zhang et al. [39]	0.758	0.608	0.437	0.711	0.465	0.546	0.555	0.549	0.424	0.358	0.541
Tsai et al. [30]	0.693	**0.761**	0.572	0.704	**0.677**	0.597	**0.642**	0.571	0.441	**0.579**	0.623
OURS	**0.811**	0.688	**0.634**	**0.738**	0.597	**0.645**	0.634	**0.582**	**0.524**	0.455	**0.631**

Table 2. Results on the DAVIS dataset in terms of IOU (higher values indicate better results).

	[20]	[28]	[31]	[2]	OURS		[20]	[28]	[31]	[2]	OURS		[20]	[28]	[31]	[2]	OURS
bear	**.898**	.864	.657	.851	.748	drtC	.667	.314	.244	**.758**	.559	motoj	.602	.245	.491	**.618**	.608
bswan	.732	.422	.223	.526	**.876**	drtS	**.683**	.344	.268	.575	.623	mbike	.559	.387	.335	**.738**	.476
bumps	.241	**.368**	.188	.353	.297	drtT	.533	.615	.349	.638	**.678**	parag	.725	.890	.568	**.933**	.726
trees	.180	.121	.194	.188	**.350**	eleph	**.824**	.494	.510	.689	.756	paral	.506	.591	.539	.512	**.628**
boat	**.361**	.056	.271	.144	.359	flamg	**.817**	.783	.570	.794	.381	park	.458	.146	.392	.295	**.677**
bdan	**.467**	.183	.422	.236	.371	goat	.554	.074	.257	**.735**	.728	rhino	.776	.520	.685	**.902**	.552
bdanF	.616	.317	.476	.157	**.700**	hike	.889	.878	.683	.603	**.893**	rolb	.318	.406	.141	**.801**	.125
bus	.825	.664	.739	**.885**	.815	hockey	.467	**.817**	.566	.713	.602	scbla	.522	**.759**	.348	.579	.588
camel	.562	**.850**	.320	.756	.762	hjH	.578	**.830**	.568	.734	.351	scgra	.325	.327	.421	.345	**.670**
carR	.808	**.872**	.500	.630	.768	hjL	.526	**.743**	.388	.682	.411	sobox	.410	**.832**	.332	.672	.578
carS	.698	.759	.538	**.880**	.781	ksurf	.272	.357	.193	.419	**.583**	socB	**.843**	.242	.378	.370	.490
carT	**.851**	.820	.611	.621	.754	kwalk	.649	.447	.724	.597	**.733**	strol	.580	.619	.466	**.678**	.654
cows	.791	.562	.623	**.799**	.770	libby	.507	.169	.470	.050	**.508**	surf	.475	.273	.312	.770	**.870**
jump	**.598**	.341	.291	.065	.342	lucia	.644	**.840**	.706	.417	.833	swing	.431	.533	.569	.622	**.755**
twirl	.453	.452	.372	.366	**.461**	malf	.601	.380	.227	.033	**.708**	tennis	.388	.494	.480	.590	**.625**
dog	.708	.753	.566	.331	**.856**	malw	.087	.245	.085	.045	**.658**	train	.831	**.903**	.620	.887	.736
agid	**.280**	.193	.055	.110	.071	motob	.617	.603	.351	.466	**.750**	Ave.	.575	.514	.426	.543	**.612**

reduced to one fifth after each iteration. The momentum and weight decay were fixed to 0.9 and 0.0005, respectively, during the entire learning iterations. The converge threshold T was set to 0.85. Notice that the parameters in weakly supervised learning methods usually cannot be tuned as there is no GT data. In this work, we set $\lambda = 0.8$ initially and then increased 0.04 after each iteration, which enables the learner selecting large part of data for training. γ was set equally to λ and τ was set to 1 to reflect the equally important of the group priority term and the diversity term.

3.2. Comparison with the State-of-the-Arts

In this section, we compared the proposed approach with the state-of-the-art methods on two benchmark datasets. Specifically, on the Youtube-Object dataset, we compared our approach with [39, 27, 31, 20, 35, 30], and on the DAVIS dataset, we compared our approach with [20, 28, 31, 2]. These compared approaches are the state-of-the-art weakly supervised or unsupervised video object segmentation approaches which are accessible for the corresponding dataset. We did not compare with the semi-supervised learning-based approaches as they need stronger supervision. For quantitative evaluation, we reported the evaluation results on the YouTube-Object dataset and DAVIS dataset in Table 1 and Table 2, specifically, which not only show the average performance in the whole datasets but also the detailed performance in each semantic category. Encouragingly, on the YouTube-Object dataset, the proposed approach can achieve promising experimental results which outperform the previous methods in most object categories as well as the average performance. On the DAVIS dataset, the proposed approach also obtains the superior average performance. Obvious performance gain can be observed in some categories like "dog" and "park". Thus, the experimental results can evidently demonstrate the effectiveness of the proposed approach. We also visualized some experimental results in Fig. 2. Failure cases might be caused by the highly confusing appearance between the foreground objects and their surrounding background regions as well as the limited training data.

3.3. Model Analysis

In this section, we first validated the effectiveness of SPFTN by comparing it with several baseline strategies as shown in Table. 3. From the experimental results, we can observe that: Segmentation proposals obtained by using object detectors as in [39] can obtain better performance due to the stronger supervision. 2) Directly using the *PTnet* cannot obtain satisfactory segmentation results as the network only encodes general object knowledge, which cannot recognize

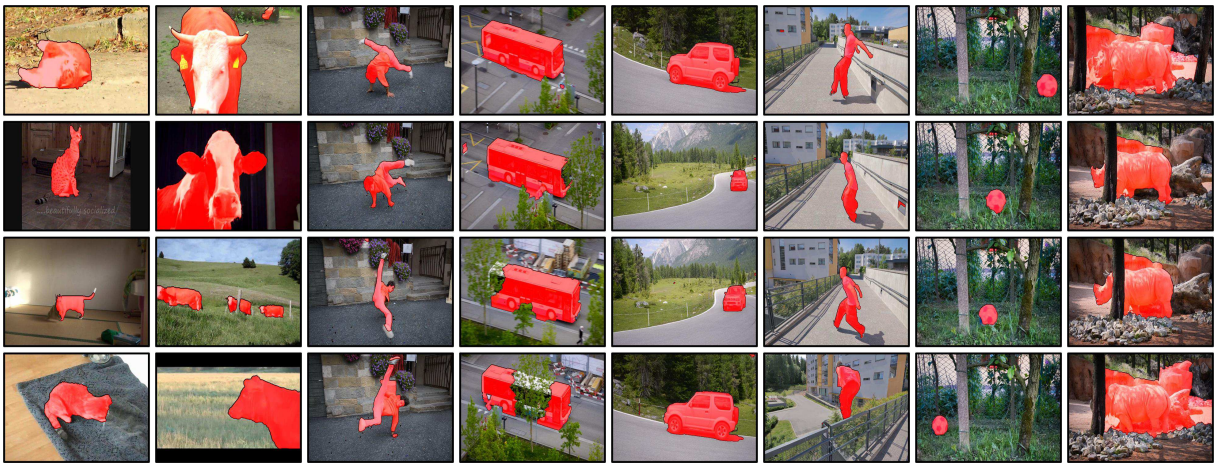

Figure 2. Some visualization examples of our experimental results. Examples in the first two columns are from the YouTube-Object dataset. Other examples are from the DAVIS dataset. The last column is the examples of failure case.

Table 3. Comparison with other baselines on YouTube-Object.

Baselines	IOU
The adopted segmentation proposal	0.510
Segmentation proposal obtained by object detectors	0.561
PTnet: pre-trained network on MSRA	0.507
Cnet: fine-tune PTnet w/o SPL	0.563
Cnet+updation: additionally update GT	0.575
Cnet-Imagenet: Cnet w/o using MSRA	0.555
OURS-Imagenet: OURS w/o using MSRA	0.602
OURS-GC: OURS w/o group curriculum	0.623
OURS	**0.631**

the specific semantics in the given videos. 3) The *Cnet* and *Cnet+updating* could only improve the performance of the *PTnet* to a limited extent or even weaken the performance like in the "car" category due to the fact that it is lack of effective learning regime to overcome the data ambiguity under the weak supervision. 4) The proposed SPFTN consistently outperforms other baselines as its learning procedure has been guided by the SPL regime. 5) The proposed group curriculum-based regularizer can effectively boost the performance as compared with the conventional self-paced regularizer. 6) Pre-training on salient object dataset could encode helpful saliency priors, which benefits learning under the weak supervision. However, without using such prior knowledge, our approach (*OUTS-Imagenet*) can still outperform most state-of-the-arts.

Then, for further demonstrating the effectiveness of the proposed self-paced regularizer, we equipped the proposed learning framework with different self-paced regularizers and compared the experimental results on the DAVIS dataset. The results reported in Table.4 indicate that each of the regularization terms used in the proposed group curriculum regularizer can benefit the learning procedure, while simultaneously using both of them obtains more significant performance gain. In addition, the proposed regularizer can also outperform the self-paced regularizer in the state-of-the-art self-paced learning approach [13].

Table 4. Evaluation of the self-paced regularizers on DAVIS.

Different regularizers	IOU
OURS-GC: OURS w/o group curriculum	0.569
OURS-GC2: OURS w/o the second term in GC	0.584
OURS-GC1: OURS w/o the first term in GC	0.589
OURS with sample diversity term of [13]	0.583
OURS	**0.612**

Another interesting experiment is to see whether the proposed method can leverage the negative samples which might be collected easily. To this end, we simply used the negative mining method [38] to help selecting initial proposals. We trained the network on aero class for ten times. The negative data are randomly sampled from other classes. The obtained performance ranges from 0.76 to 0.83, which could improve our result (0.81) but not stable just as we analyzed in previous sections.

4. Conclusion

This paper has proposed a novel SPFTN-based framework for segmenting objects in weakly labelled videos. By integrating the self-paced learning regime and the learning function of the DNN into a unified and compatible framework, the proposed approach can effectively fine-tune DNN under weak supervision. Comprehensive experiments on the large-scale YouTube-Object and DAVIS datasets have demonstrated the effectiveness of both the entire SPFTN framework and the newly proposed group curriculum regularizer. In future, we plan to further improve the learning regime and apply it in other weakly supervised learning tasks like weakly supervised image segmentation [23, 21, 14] and co-saliency detection [34].

Acknowledgement: This work was supported in part by the National Science Foundation of China under Grant 61473231, Grant 61522207, Grant 61373114, and Grant 61661166011.

References

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[2] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[3] X. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016.

[4] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li. Disc: Deep image saliency computing via progressive representation learning. *TNNLS*, 27(6):1135–1149, 2016.

[5] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2015.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[7] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu. Background prior-based salient object detection via deep reconstruction residual. *TCSVT*, 25(8):1309–1321, 2015.

[8] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV*, 2012.

[9] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *ECCV*, 2014.

[10] K. R. Jerripothula, J. Cai, and J. Yuan. Cats: Co-saliency activated tracklet selection for video co-localization. In *ECCV*, 2016.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM-MM*, 2014.

[12] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM-MM*, 2014.

[13] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014.

[14] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[15] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[16] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[17] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, 2015.

[18] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu. Weakly supervised multiclass video segmentation. In *CVPR*, 2014.

[19] D. Meng, Q. Zhao, and L. Jiang. What objective does self-paced learning indeed optimize? *arXiv preprint arXiv:1511.06049*, 2015.

[20] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[21] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[22] F. Perazzi, J. P.-T. B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[23] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[24] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012.

[27] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013.

[28] B. Taylor, V. Karasev, and S. Soattoc. Causal video object segmentation from persistence of occlusions. In *CVPR*, 2015.

[29] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[30] Y.-H. Tsai, G. Zhong, and M.-H. Yang. Semantic co-segmentation in videos. In *ECCV*, 2016.

[31] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015.

[32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[33] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang. Revealing event saliency in unconstrained video collection. *TIP*, 26(4):1746–1758, 2017.

[34] D. Zhang, J. Han, C. Li, J. Wang, and X. Li. Detection of co-salient objects by looking deep and wide. *IJCV*, 120(2):215–232, 2016.

[35] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013.

[36] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han. A self-paced multiple-instance learning framework for co-saliency detection. In *ICCV*, 2015.

[37] D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In *IJCAI*, 2016.

[38] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *CVPR*, 2015.

[39] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia. Semantic object segmentation via detection in weakly labeled video. In *CVPR*, 2015.

[40] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, 2015.