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] and Junwei Han[1]

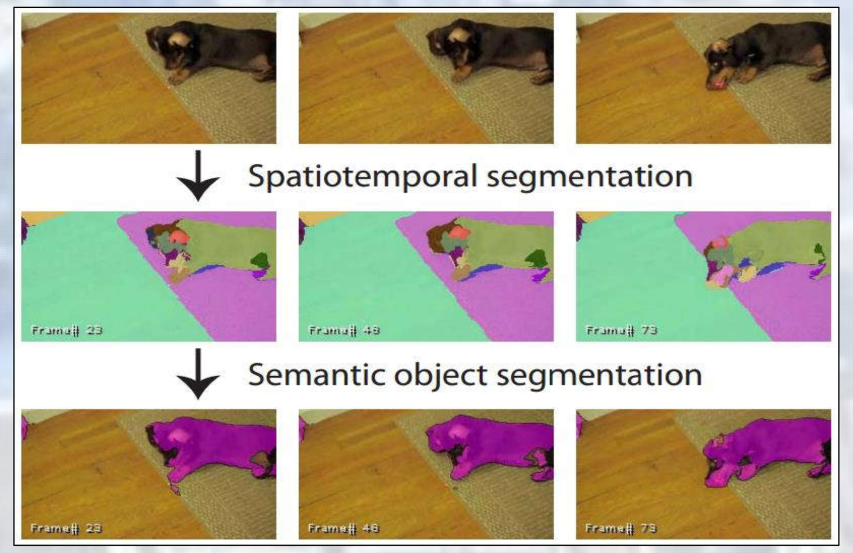
[1]Northwestern Polytechnical University, [2]Xi'an Jiaotong University, [3]University of Sydney

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR 2017, July 21-26, Honolulu

Problem

Goal: learning to perform category-specific video object segmentation by only using video-level tags.



from Tang's CVPR 2013

Challenges:
Detecting!
Associating!
Recognizing!
Segmenting!

Conventional approaches:

- Decompose positive and negative videos into **spatial-temporal segments**.
- Train segmentation-level **classifiers or inference models** under the weak supervision.
- **Identify** the segments related to the given object categories in each video.

Under studied problems:

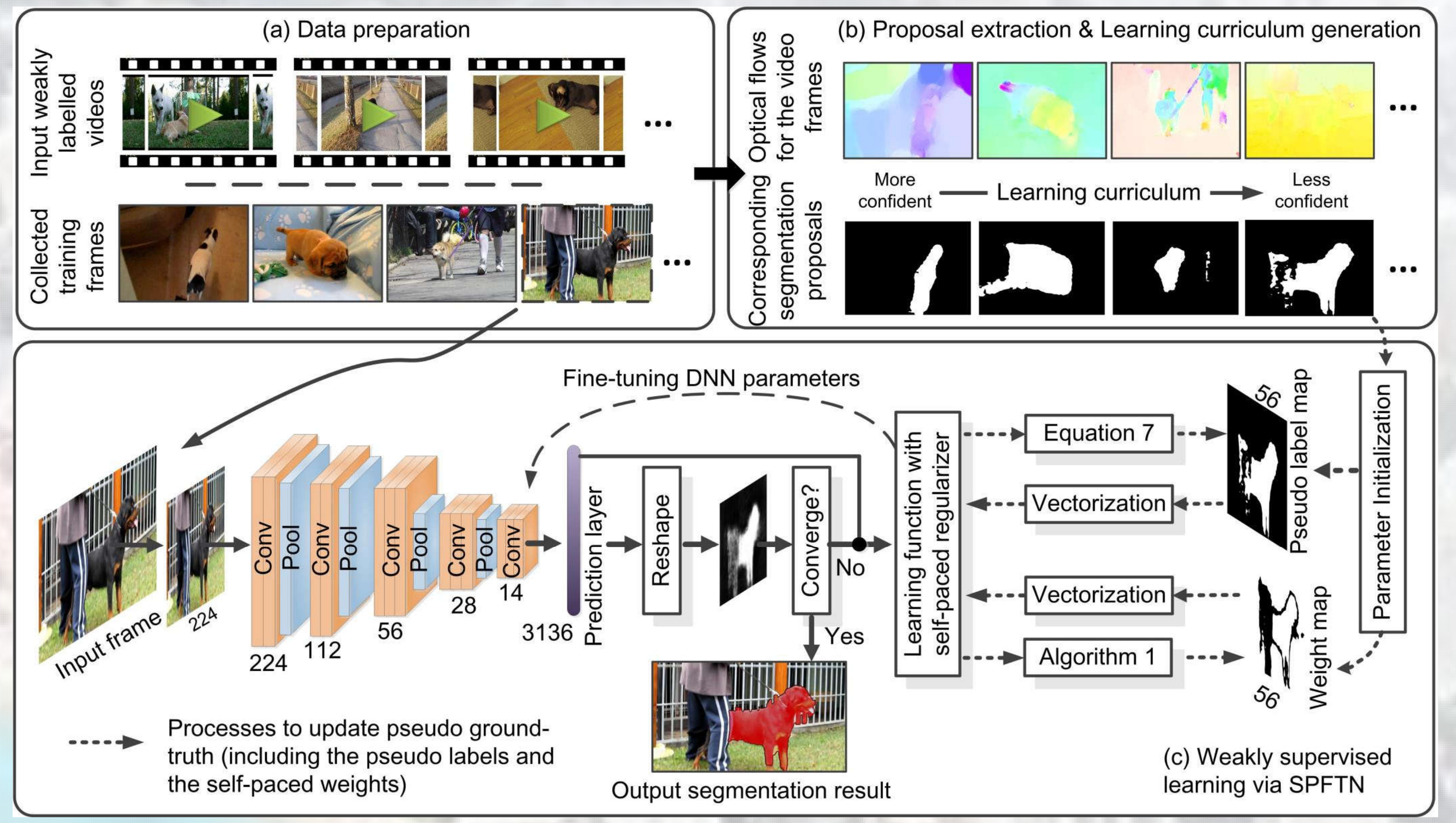
- Unclear how to address this problem via leveraging powerful **DNNs**.
- Explore **scene context** in each video frame rather than consider each spatial-temporal segment individually to provide helpful contextual priors.
- Alleviate the **learning uncertainty** brought by the **negative videos** due to the lack of principle ways to acquire them.

Solution: SPFTN!

Our Approach

Main Idea:

- Integrate **SPL** into the **DNN** learning objective to improve the learning capability of SPL and simultaneously perform **weakly supervised training of DNN**.
- Use a novel **group curriculum** self-paced term to encode helpful prior-knowledge.
- Capture object semantics **only** from **positive videos** to increase **learning stability**.
- Encode **rich context** information to help improve the segmentation accuracy.



Learning Objective:

$$\min_{\mathbf{W},\mathbf{Y},\mathbf{V}} \mathbf{E}(\mathbf{W},\mathbf{Y},\mathbf{V}) = r(\mathbf{W}) + \sum_{k=1}^{K} L(\mathbf{y}_k, \mathbf{v}_k, \Phi(I_k|\mathbf{W})) + f(\mathbf{V}; \mathbf{p}, \lambda, \gamma, \tau),$$

$$s.t. \sum_{k} \|\mathbf{v}_k\|_1 \in (0, d \times K), \quad \sum_{k} \|\mathbf{y}_k\|_1 \in (0, d \times K).$$

Self-paced Regularizer:

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

$$-\gamma \left(\underbrace{\sum_{k=1}^{K} p_k \sqrt{\sum_{i=1}^{d} v_k^i}}_{\text{Group priority}} + \tau \underbrace{\sum_{k=1}^{K} \sqrt{\sum_{i=1}^{d} v_k^i}}_{\text{Group diversity}} \right).$$

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

input : Videos weakly labelled as containing a certain type of object and the pre-trained network;
output: The semantic object segmentation masks for each video frame;

```
Collect video frames and the corresponding segmentation proposals with data augmentation;
Obtain learning curriculum by calculating p_k;
Initialize the pseudo labels Y, the self-paced weights V, and assign the parameter values λ, γ, and τ;
while not converge do
    Fine-tune the DNN parameters W via Eq. (5);
    Update the pseudo labels Y via Eq. (6);
    Update the self-paced weights 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.
```

Experiments

Dataset: YouTube-Object dataset & DAVIS

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**

State-of-the-arts:

Ablation Study:

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**

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**