Two-shot Video Object Segmentation

Kun Yan\textsuperscript{1} Xiao Li\textsuperscript{2} Fangyun Wei\textsuperscript{2} Jinglu Wang\textsuperscript{2} Chenbin Zhang\textsuperscript{1} Ping Wang\textsuperscript{1*} Yan Lu\textsuperscript{2*}

\textsuperscript{1}Peking University \textsuperscript{2}Microsoft Research Asia

{kyan2018, zcb, pfw}@pku.edu.cn \{xili11, fawe, jinglwa, yanlu\}@microsoft.com

Abstract

Previous works on video object segmentation (VOS) are trained on densely annotated videos. Nevertheless, acquiring annotations in pixel level is expensive and time-consuming. In this work, we demonstrate the feasibility of training a satisfactory VOS model on sparsely annotated videos—we merely require two labeled frames per training video while the performance is sustained. We term this novel training paradigm as two-shot video object segmentation, or two-shot VOS for short. The underlying idea is to generate pseudo labels for unlabeled frames during training and to optimize the model on the combination of labeled and pseudo-labeled data. Our approach is extremely simple and can be applied to a majority of existing frameworks. We first pre-train a VOS model on sparsely annotated videos in a semi-supervised manner, with the first frame always being a labeled one. Then, we adopt the pre-trained VOS model to generate pseudo labels for all unlabeled frames, which are subsequently stored in a pseudo-label bank. Finally, we retrain a VOS model on both labeled and pseudo-labeled data without any restrictions on the first frame. For the first time, we present a general way to train VOS models on two-shot VOS datasets. By using 7.3\% and 2.9\% labeled data of YouTube-VOS and DAVIS benchmarks, our approach achieves comparable results in contrast to the counterparts trained on fully labeled set. Code and models are available at https://github.com/ykpku/Two-shot-Video-Object-Segmentation.

1. Introduction

Video object segmentation (VOS), also known as mask tracking, aims to segment the target object in a video given the annotation of the reference (or first) frame. Existing approaches\cite{7, 9, 21, 30, 37, 46, 52} are trained on densely annotated datasets such as DAVIS\cite{33, 34} and YouTube-VOS\cite{50}. However, acquiring dense annotations, particularly at the pixel level, is laborious and time-consuming. For instance, the DAVIS benchmark consists of 60 videos, each with an average of 70 labeled frames; the YouTube-VOS dataset has an even larger amount of videos, and every fifth frame of each video is labeled to lower the annotation cost. It is necessary to develop data-efficient VOS models to reduce the dependency on labeled data.

In this work, we investigate the feasibility of training a
satisfactory VOS model on sparsely annotated videos. For the sake of convenience, we use the term $N$-shot to denote that $N$ frames are annotated per training video. Note that 1-shot is meaningless since it degrades VOS to the task of image-level segmentation. We use STCN [9] as our baseline due to its simplicity and popularity. Since at least two labeled frames per video are required for VOS training, we follow the common practice to optimize a naive 2-shot STCN model on the combination of YouTube-VOS and DAVIS, and evaluate on YouTube-VOS 2018/2019 and DAVIS 2016/2017, respectively. We compare the native 2-shot STCN with its counterpart trained on full set in Fig. 1b. Surprisingly, 2-shot STCN still achieves decent results, for instance, only a $-2.1\%$ performance drop is observed on YouTube-VOS 2019 benchmark, demonstrating the practicality of 2-shot VOS.

So far, the wealth of information present in unlabeled frames is yet underexplored. In the last decades, semi-supervised learning, which combines a small amount of labeled data with a large collection of unlabeled data during training, has achieved considerable success on various tasks such as image classification [3, 39], object detection [40, 49] and semantic segmentation [14, 17]. In this work, we also adopt this learning paradigm to promote 2-shot VOS (see Fig. 1a). The underlying idea is to generate credible pseudo labels for unlabeled frames during training and to optimize the model on the combination of labeled and pseudo-labeled data. Here we continue to use STCN [9] as an example to illustrate our design principle, nevertheless, our approach is compatible with most VOS models. Concretely, STCN takes a randomly selected triplet of labeled frames as input but the supervisions are only applied to the last two—VOS requires the annotation of the first frame as reference to segment the object of interest that appeared in subsequent frames. This motivates us to utilize the ground-truth for the first frame to avoid error propagation during early training. Each of the last two frames, nevertheless, can be either a labeled frame or an unlabeled frame with a high-quality pseudo label. Although the performance is improved with this straightforward paradigm, the capability of semi-supervised learning is still underexplored due to the restriction of employing the ground truth as the starting frame. We term the process described above as phase-1.

To take full advantage of unlabeled data, we lift the restriction placed on the starting frame, allowing it to be either a labeled or pseudo-labeled frame. To be specific, we adopt the VOS model trained in phase-1 to infer the unlabeled frames for pseudo-labeling. After that, each frame is associated with a pseudo label that approximates the ground-truth. The generated pseudo labels are stored in a pseudo-label bank for the convenience of access. The VOS model is then retrained without any restrictions—similar to how it is trained through supervised learning, but each frame has either a ground-truth or a pseudo-label attached to it. It is worth noting that, as training progresses, the predictions become more precise, yielding more reliable pseudo labels—we update the pseudo-label bank once we identify such pseudo labels. The above described process is named as phase-2. As shown in Fig. 1b, our approach assembled onto STCN, achieves comparable results (e.g. 85.2\% v.s 85.1\% on DAVIS 2017, and 82.7\% v.s 82.7\% on YouTube-VOS 2019) in contrast to its counterpart, STCN trained on full set, though our approach merely accesses 7.3\% and 2.9\% labeled data of YouTube-VOS and DAVIS benchmark, respectively.

Our contributions can be summarized as follows:

- For the first time, we demonstrate the feasibility of two-shot video object segmentation: two labeled frames per video are almost sufficient for training a decent VOS model, even without the use of unlabeled data.
- We present a simple yet efficient training paradigm to exploit the wealth of information present in unlabeled frames. This novel paradigm can be seamlessly applied to various VOS models, e.g., STCN [9], RDE-VOS [21] and XMem [7] in our experiments.
- Though we only access a small amount of labeled data (e.g. 7.3\% for YouTube-VOS and 2.9\% for DAVIS), our approach still achieves competitive results in contrast to the counterparts trained on full set. For example, 2-shot STCN equipped with our approach achieves 85.1\%/82.7\% on DAVIS 2017/YouTube-VOS 2019, which is +4.1\%/+2.1\% higher than the naive 2-shot STCN while -0.1\%/-0.0\% lower than the STCN trained on full set.

2. Related work

**Video object segmentation.** Existing VOS methods can be categorized into two groups: online-learning methods and offline-learning methods. Online-learning methods [4, 10, 24, 25, 32, 42, 45] need to fine-tune the networks at test time based on the query mask of the first frame. However, test-time fine-tuning is computationally expensive. In contrast, offline-learning methods [12, 15, 23, 26, 53, 55] aim at training a model that segments videos without any adaptations during inference. It is usually achieved via propagation and matching. Propagation-based methods [5, 16, 22, 29] segment the target object sequentially by propagating the reference mask of the first frame. Matching-based methods [9, 30, 43, 52] typically employ a memory bank to store the features of a collection of frames, then a feature matching is adopted to segment the query frame.
STM [30] received widespread attention among the matching-based methods. STM proposes to construct a memory network to store the masks of the previous frames. Then the query frame is segmented using the information stored in the memory. A majority of follow-up works improved STM in several aspects [7, 21, 36, 37, 46]. For example, STCN [9] establishes correspondences between frames to avoid re-encoding the mask feature of each object; RDE-VOS [21] builds a constant-size memory bank by recurrent dynamic embedding while retaining the performance; XMem [7] incorporates multiple feature memory stores and achieves the best performance. Despite their promising results, these methods need densely annotated videos for training. Instead, our method only needs two labeled frames per video and is compatible with most VOS models. It is worth noting that the meaning of “one-shot” claimed by [4] significantly differs from that of our “two-shot”. In [4], “one-shot” refers to that given a reference frame during inference, the optimized model is able to segment the remaining frames. In contrast, we use the term “N-shot” to denote the number of labeled frames per video. Therefore, in our setting, “one-shot” denotes that only a single labeled frame per video is available during training.

**Semi-supervised learning.** Semi-supervised learning is an efficient way to improve model performance by using a few labeled data and a large amount of unlabeled data. It has achieved promising results across various computer vision tasks, such as image classification [39, 41], image segmentation [14, 17], object detection [40, 49] and action recognition [51]. The dominated works can be roughly categorized into consistency based methods [3, 6, 11, 18, 35, 41] and pseudo-labeling based methods [13, 19, 40, 48, 56, 57]. Consistency based methods enforce consistency between predictions of different perturbations, such as model perturbing [1], data augmentations [2, 47] and adversarial perturbations [28]. Pseudo-labeling based methods generate one-hot pseudo labels for unlabeled data. Then the model is optimized on the combination of labeled data and pseudo-labeled data. Our approach also adopts pseudo-labeling to improve two-shot VOS.

## 3. Methodology

We first revisit the preliminary of VOS in Section 3.1. Then we formulate the problem of two-shot VOS and show an overview of our method in Section 3.2. Next, the details of training a two-shot VOS model are presented in Section 3.3 and 3.4. At last, we show our methodology can be generalized to a majority of VOS models in Section 3.5.

### 3.1. Preliminary

Previous works train VOS models on densely annotated videos. Given the annotation of the first frame, the training objective is to maximize the mask prediction of the target object from the second frame to the last frame. For instance, STM [30] and STCN [9] take a triplet of frames as input; RDE-VOS [21] and XMem [7] propose to model longer video sequences containing 5 and 8 frames, respectively. Random frame skipping, which randomly skips frames during the sampling, is a widely-used data augmentation to improve the generalization. In general, the maximum number of frames to skip gradually increases from 0 to K as training progresses.

In our setting, we could only access two labeled frames per video. To reduce error propagation caused by unreliable pseudo labels, we adopt STCN [9] as our base model in phase-1 training since it merely needs a triplet of frames as input. Nevertheless, we can train any VOS models in phase-2, which will be described in Section 3.5. Now we briefly revisit STCN. Given a training video, STCN first samples a triplet of frames as input. Then it predicts the mask of the second frame according to the ground-truth of the first frame, and the mask of the third frame based on the prediction of the previous frame in addition to the ground-truth of the first frame. The objective function of STCN is a standard segmentation loss, which is applied to each of the two predictions.

### 3.2. Problem formulation and overview

**Problem formulation.** Given a VOS dataset \( \mathcal{D} \), for each training video \( \mathcal{V} = [V_1, ..., V_T] \in \mathcal{D} \) containing \( T \) \( (T \gg 2) \) frames with the associated ground-truth \( \mathcal{Y} = [Y_1, ..., Y_T] \), we randomly sample two frames as the labeled data, while the remaining ones are served as the unlabeled data. The objective is to train a VOS model by using both labeled and unlabeled data.

**Overview.** Fig. 2 shows an overview of our two-shot video object segmentation (VOS). First, we train a VOS model in a semi-supervised manner, with the reference frame always being a labeled one, which is referred to as phase-1 training. Then, we perform an intermediate inference to generate pseudo labels for unlabeled frames by the VOS model trained in phase-1. The generated pseudo labels are stored in a pseudo-label bank for the convenience of accessing. At last, we re-train a VOS model on both labeled frames and pseudo-labeled frames without any restrictions on the reference frame. We term this stage as phase-2 training. It is worth noting that the pseudo-label bank is dynamically updated once more reliable pseudo labels are yielded in phase-2 training.

### 3.3. Phase-1 training

We adopt STCN [9] as our base model, which takes a triplet of frames as input. Nevertheless, in our setting, each training video only contains two labeled frames, which is insufficient to be served as the input of STCN in a fully supervised manner. To tackle this problem, we adopt
semi-supervised learning, which generates pseudo-labeled frames together with the labeled ones to enable triplet construction. Since STCN requires the annotation of the reference (or first) frame to segment the object of interest that appeared in subsequent frames, we always use a labeled frame as the reference frame to alleviate the error propagation in the phase-1 training. The last two frames, however, can be either labeled or unlabeled. In our implementation, the last two frames have a 0.5 probability of being both unlabeled, and a 0.5 probability of having one frame be labeled. The training of two-shot VOS is identical to that of full-set VOS, except that our training triplet is composed of labeled frames with ground-truth and unlabeled frames with pseudo labels. Concretely, given a randomly sampled triplet where the last two frames are composed of $N_1$ labeled frames and $N_2$ unlabeled frames ($N_1 = 1, N_2 = 1$ or $N_1 = 0, N_2 = 2$), the overall loss $\mathcal{L}$ is the sum of the supervised loss $\mathcal{L}_S$ and the unsupervised loss $\mathcal{L}_U$ defined as follows:

$$
\mathcal{L}_S = \frac{1}{HWN_1} \sum_{n=1}^{N_1} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{H}(Y_n^{(i,j)}, P_n^{(i,j)}),
$$

$$
\mathcal{L}_U = \frac{1}{HWN_2} \sum_{n=1}^{N_2} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbb{1}_{[\max(P_n^{(i,j)}) \geq \tau_1]} \mathcal{H}(Y_n^{(i,j)}, \hat{P}_n^{(i,j)}),
$$

where $H$ and $W$ represent the height and the width of the input, $\mathcal{H}(...)$ denotes the cross-entropy function, $P_n^{(i,j)}$ is the prediction at pixel $(i,j)$ in the $n$-th labeled frame, and $Y_n^{(i,j)}$ denotes the corresponding ground-truth.

The unsupervised loss $\mathcal{L}_U$ is a variant of $\mathcal{L}_S$, which is defined as follows:

$$
\mathcal{L}_U = \frac{1}{HWN_2} \sum_{n=1}^{N_2} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbb{1}_{[\max(P_n^{(i,j)}) \geq \tau_1]} \mathcal{H}(Y_n^{(i,j)}, \hat{P}_n^{(i,j)}),
$$

where $\mathbb{1}_{[\cdot]}$ is the indicator function to filter out the predictions whose maximal confidences are lower than the predefined threshold $\tau_1$, $\hat{P}_n^{(i,j)}$ is the prediction at pixel $(i,j)$ in the $n$-th unlabeled frame, and $\hat{Y}_n^{(i,j)} = \text{argmax}(P_n^{(i,j)})$ represents the corresponding one-hot pseudo label. By default, we set $\tau_1 = 0.9$ to guarantee the reliability of the yielded pseudo labels.

As training progresses, an increasing number of high-quality pseudo-labeled samples are generated, injecting implicit knowledge included in unlabeled data into the model. In addition, we also randomly skip frames during sampling as described in Section 3.1.
The training process of phase-2 is identical to training. The inference of a VOS model requires the annotation of the reference frame, allowing it to be either a labeled or pseudo-labeled frame. The underlying idea behind phase-2 training is to generate pseudo labels for all unlabeled frames using the decent VOS model trained in phase-1. After then, the pseudo-labeled data is stored in a pseudo-label bank, providing efficient access when constructing a training triplet where the reference frame is selected as a pseudo-labeled one.

Intermediate inference and pseudo-label bank. We perform an intermediate inference before initiating phase-2 training. The inference of a VOS model requires the annotation of the reference frame. Nonetheless, only two labeled frames per video are available in our scenario. To generate the pseudo label per frame, inspired by bidirectional prediction and labelling [20, 27], we introduce a bidirectional inference strategy as shown in Figure 3. Specifically, for each of the two labeled frames, the VOS model trained in phase-1 takes it as the reference frame to infer the predictions for the unlabeled frames from the inference frame to the end frame and, in a reverse manner, from the inference frame to the beginning frame. After that, each unlabeled frame has two predictions associated with it, and we pick the prediction inferred by the labeled frame that is closest to this unlabeled frame. We maintain a pseudo-label bank to store pseudo labels associated with unlabeled frames.

Training. The training process of phase-2 is identical to that of phase-1, except that the reference (or first) frame can be either a labeled frame or an unlabeled frame with a pseudo label from the pseudo label bank attached to it.

**Update pseudo-label bank.** As training progresses, predictions become more accurate, resulting in more reliable pseudo labels. Therefore, to further facilitate phase-2 training, we propose to dynamically update the pseudo-label bank as needed. Concretely, at each iteration, given the prediction $P$ of an unlabeled frame, we use $P^{(i,j)}$ to denote the prediction at pixel $(i,j)$. Once the prediction $P^{(i,j)}$ meets the condition that $\max(P^{(i,j)}) \geq \tau_2$, where $\tau_2$ denotes a pre-defined threshold, the corresponding pseudo label in pseudo label bank is updated by $\hat{Y}^{(i,j)} = \text{argmax}(P^{(i,j)})$. We set $\tau_2 = 0.99$ by default.

3.5. Generalization capability

Thanks to the proposed pseudo-label bank and phase-2 training, our two-shot training paradigm can be applied to a majority of VOS models regardless of their architectures and requirements on the input. To generalize to other models, we adopt a STCN model trained in phase-1 to construct a pseudo-label bank. After that, various VOS models can utilize the universal training paradigm presented in phase-2 to enable two-shot VOS learning. Experimentally, we also apply our methodology to RDE-VOS [21] and XMem [7] besides STCN [9] to show the generalization capability.

4. Experiments

4.1. Experimental setup

**Datasets.** We conduct experiments on widely used VOS benchmarks including DAVIS 2016/2017 [33, 34] and YouTube-VOS 2018/2019 [50]. DAVIS 2017 is a multi-object extension of DAVIS 2016, which consists of 60 (138 objects) and 30 (59 objects) videos for training and validation respectively. YouTube-VOS is a larger-scale multi-object dataset with 3471 videos from 65 categories for training. These training videos are annotated every five frames. There are 474 and 507 videos in the 2018 and 2019 validation splits respectively. In our two-shot setting, we randomly select two labeled frames per video as labeled data while the remaining ones are served as unlabeled data. Compared to full set, we only use 7.3% and 2.9% labeled data for YouTube-VOS and DAVIS, respectively.

**Evaluation metric.** Following common practice [7, 9, 30], for the DAVIS datasets, we adopt the standard metrics: region similarity $J$, contour accuracy $F$ and their average $JF$. For the YouTube-VOS datasets, we report $J$ and $F$ of the seen and unseen categories, and their averaged score $G$.

**Implementation details.** We implement our method with PyTorch [31]. For phase-1 training, we adopt the STCN [9] pre-trained on static image datasets [38, 44, 54] with synthetic deformations. The parameter $K$ in random frame...
Table 1. Comparison with different methods on YouTube-VOS 2018 and 2019 validation sets. Subscripts S and U denote seen and unseen categories respectively. * denotes reproduced result by using the open-source code. By using 7.3% labeled data (2 labeled frames per training video) of YouTube-VOS benchmark, our approach achieves comparable results in contrast to the counterpart trained on full set, and outperforms the native 2-shot counterpart by large margins.

<table>
<thead>
<tr>
<th>Method</th>
<th>Labeled data</th>
<th>YouTube-VOS 2018</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G</td>
<td>J</td>
<td>F</td>
<td>J</td>
<td>F</td>
<td>J</td>
<td>F</td>
<td>J</td>
<td>F</td>
</tr>
<tr>
<td>STM [30]</td>
<td>100%</td>
<td>79.4</td>
<td>79.7</td>
<td>84.2</td>
<td>72.8</td>
<td>80.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MiVOS [8]</td>
<td>100%</td>
<td>80.4</td>
<td>80.0</td>
<td>84.6</td>
<td>74.8</td>
<td>82.4</td>
<td>80.3</td>
<td>79.3</td>
<td>83.7</td>
<td>75.3</td>
</tr>
<tr>
<td>CFBI [52]</td>
<td>100%</td>
<td>81.4</td>
<td>81.1</td>
<td>85.8</td>
<td>75.3</td>
<td>83.4</td>
<td>81.0</td>
<td>80.6</td>
<td>85.1</td>
<td>75.2</td>
</tr>
<tr>
<td>RDE-VOS [21]</td>
<td>100%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>81.9</td>
<td>81.1</td>
<td>85.5</td>
<td>76.2</td>
</tr>
<tr>
<td>HMMN [37]</td>
<td>100%</td>
<td>82.6</td>
<td>82.1</td>
<td>87.0</td>
<td>76.8</td>
<td>84.6</td>
<td>82.5</td>
<td>81.7</td>
<td>86.1</td>
<td>77.3</td>
</tr>
<tr>
<td>JOINT [26]</td>
<td>100%</td>
<td>83.1</td>
<td>81.5</td>
<td>85.9</td>
<td>78.7</td>
<td>86.5</td>
<td>82.7</td>
<td>81.1</td>
<td>85.4</td>
<td>78.2</td>
</tr>
<tr>
<td>STCN [9]</td>
<td>100%</td>
<td>83.0</td>
<td>81.9</td>
<td>86.5</td>
<td>77.9</td>
<td>85.7</td>
<td>82.7</td>
<td>81.1</td>
<td>85.4</td>
<td>78.2</td>
</tr>
<tr>
<td>R50-AOT-L [53]</td>
<td>100%</td>
<td>84.1</td>
<td>83.7</td>
<td>88.5</td>
<td>78.1</td>
<td>86.1</td>
<td>84.1</td>
<td>83.5</td>
<td>88.1</td>
<td>78.4</td>
</tr>
<tr>
<td>XMem [7]</td>
<td>100%</td>
<td>85.7</td>
<td>84.6</td>
<td>89.3</td>
<td>80.2</td>
<td>88.7</td>
<td>85.5</td>
<td>84.3</td>
<td>88.6</td>
<td>80.3</td>
</tr>
</tbody>
</table>

Table 2. Comparisons with different methods on DA VIS benchmark. Our approach achieves comparable results in contrast to the counterparts trained on full sets; (2) their counterparts trained on two-shot datasets without using unlabeled data; skipping is gradually increased from 5 to 25 with a curriculum learning schedule. The threshold \( \tau_1 \) is set to 0.9. The training paradigm of two-shot VOS can be seamlessly applied to various VOS models in phase-2 training. We explore STCN [9], RDE-VOS [21] and XMem [7], respectively. The threshold \( \tau_2 \) is set to 0.99.

<table>
<thead>
<tr>
<th>Method</th>
<th>Labeled data</th>
<th>DAVIS 2016</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>J &amp; F</td>
<td>J</td>
<td>F</td>
<td>J &amp; F</td>
<td>J</td>
<td>F</td>
<td>J &amp; F</td>
<td>J</td>
<td>F</td>
</tr>
<tr>
<td>STM [30]</td>
<td>100%</td>
<td>89.3</td>
<td>88.7</td>
<td>89.9</td>
<td>81.8</td>
<td>78.2</td>
<td>84.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFBI [52]</td>
<td>100%</td>
<td>89.4</td>
<td>88.3</td>
<td>90.5</td>
<td>81.9</td>
<td>79.1</td>
<td>84.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOINT [26]</td>
<td>100%</td>
<td>79.2</td>
<td>77.5</td>
<td>81.4</td>
<td>79.6</td>
<td>77.6</td>
<td>81.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDE-VOS [21]</td>
<td>100%</td>
<td>91.1</td>
<td>89.7</td>
<td>92.5</td>
<td>84.2</td>
<td>80.8</td>
<td>87.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MiVOS [8]</td>
<td>100%</td>
<td>91.0</td>
<td>89.6</td>
<td>92.4</td>
<td>84.5</td>
<td>81.7</td>
<td>87.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMMN [37]</td>
<td>100%</td>
<td>90.8</td>
<td>89.6</td>
<td>92.0</td>
<td>84.7</td>
<td>81.9</td>
<td>87.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R50-AOT-L [53]</td>
<td>100%</td>
<td>91.1</td>
<td>90.1</td>
<td>92.1</td>
<td>84.9</td>
<td>82.3</td>
<td>87.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STCN [9]</td>
<td>100%</td>
<td>91.6</td>
<td>90.8</td>
<td>92.5</td>
<td>85.4</td>
<td>82.2</td>
<td>88.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XMem [7]</td>
<td>100%</td>
<td>91.5</td>
<td>90.4</td>
<td>92.7</td>
<td>86.2</td>
<td>82.9</td>
<td>89.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2. Main results

We apply our two-shot VOS to STCN [9], RDE-VOS [21], and XMem [7], and compare the results with 1): their counterparts trained on full sets; (2) their counterparts trained on two-shot datasets without using unlabeled data;
more accurate than that in phase-1. By default, we set threshold in phase-2 training since the predictions in phase-2 are pseudo-labeling in phase-1 and -2, respectively. We adopt a higher threshold in phase-2 training since the predictions in phase-2 are more accurate than that in phase-1. By default, we set $\tau_1 = 0.9$ and $\tau_2 = 0.99$.

(3) other strong baselines trained on full sets. When training a naive 2-shot model in a fully supervised manner, we repeatedly sample the labeled frames to meet the input requirement of that model. We report the results on YouTube-VOS and DAVIS validation sets in Tab. 1 and Tab. 2, respectively. From the Tables, we could draw two conclusions: (1) Two labeled frames per video are almost sufficient for training a pleasant VOS model—even the unlabeled data are unused. For example, 2-shot STCN already achieves 80.8% score on YouTube-VOS 2018 benchmark, which is only 2.2% lower than the full-set STCN achieving 83.0% score. (2) By using 7.3% and 2.9% labeled data of YouTube-VOS and DAVIS benchmarks, our approach achieves comparable results in contrast to the counterpart trained on full set, and outperforms the native 2-shot counterpart by large margins. For instance, 2-shot STCN equipped with our approach achieves 85.1%/82.7% on DAVIS 2017/YouTube-VOS 2019, which is +4.1%/+2.1% higher than the naive 2-shot STCN while -0.1%/-0.0% lower than the full-set STCN.

4.3. Ablation study

In this section, we validate the proposed two-shot VOS training strategy step-by-step. All ablation studies are conducted on YouTube-VOS 2019 by applying our approach to STCN [9]. More analysis can be found in our supplementary material.

Effects of each phase. The results are shown in Tab. 3. Starting from a naive 2-shot STCN (denoted as "baseline" afterward) which achieves 80.6% score, phase-1 training improves the score to 81.6%. On top of this, phase-2 training further enhances performance to 82.7%, leading to the same performance of STCN trained on fully labeled set.

Thresholds of pseudo-labeling. There are two hyperparameters $\tau_1$ and $\tau_2$ controlling pseudo-labeling in phase-1 and -2, respectively. Fig. 4 displays two accuracy curves by varying $\tau_1$ and $\tau_2$. Using a higher threshold guarantees the quality of generated pseudo labels but yields less amount of pseudo data, and vice versa. We adopt a higher threshold in phase-2 training since the predictions in phase-2 are more accurate than that in phase-1. It can be seen that $\tau_1 = 0.9$ and $\tau_2 = 0.99$ yield the best result.

Different pseudo labelers. Tab. 4 ablates the effects of using different pseudo-labelers in phase-1. Specifically, we propose two variants: (1) STCN model itself; (2) STCN with a mean teacher [41] strategy. The underlying idea behind Mean Teacher (MT) is that using an exponential moving average (EMA) strategy to update the parameters of the model at each iteration, which can be formulated as: $\theta_i^t = \alpha \theta_i^{t-1} + (1 - \alpha) \theta_i$, where $t$ denotes the current iteration, $\theta_i^{t-1}$ and $\theta_i$ denote the parameters of MT-STCN and STCN respectively, and $\alpha$ is a weight. It can be seen that using the MT-STCN model surpasses the one without MT strategy. We further ablate $\alpha$ in Tab. 5. We find that $\alpha = 0.995$ yields the best performance. However, we do not employ MT strategy in phase-2 since no performance improvement is observed.

Bidirectional inference. We adopt an intermediate infer-

---

Table 3. Ablation study on the effectiveness of each phase. The naive 2-shot STCN is adopted as the baseline.

<table>
<thead>
<tr>
<th>Components</th>
<th>YouTube-VOS 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>$\mathcal{G}$</td>
</tr>
<tr>
<td>+phase-1</td>
<td>81.6+1.0</td>
</tr>
<tr>
<td>+phase-2</td>
<td>82.7+1.1</td>
</tr>
</tbody>
</table>

Table 4. Ablation study of different pseudo-labelers in phase-1. MT-STCN: the parameters of STCN are updated by a Mean Teacher [41] strategy.

<table>
<thead>
<tr>
<th>Pseudo-labeler</th>
<th>YouTube-VOS 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>$\mathcal{G}$</td>
</tr>
<tr>
<td>STCN</td>
<td>81.2+0.6</td>
</tr>
<tr>
<td>MT-STCN</td>
<td>81.6+0.4</td>
</tr>
</tbody>
</table>

Table 5. Study of different coefficient $\alpha$ used in the MT-STCN.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>YouTube-VOS 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.990</td>
<td>$\mathcal{G}$</td>
</tr>
<tr>
<td>0.995</td>
<td>81.6</td>
</tr>
<tr>
<td>0.999</td>
<td>81.3</td>
</tr>
</tbody>
</table>
formance is already saturated for two-shot VOS and acquires that of 2-shot STCN equipped with our method—the performance of 82.7% and 82.1% for 2-shot STCN and conduct one round of phase-2 training. Both models achieve the performance of 82.0% from the full set. We use PCA to visualize the feature space of naive 2-shot STCN, 2-shot STCN equipped with our approach, and full-set STCN. Orange: foreground; blue: background.

Table 6. Comparison between unidirectional inference and bidirectional inference (default).

<table>
<thead>
<tr>
<th>Update</th>
<th>YouTube-VOS 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mathcal{G} )</td>
</tr>
<tr>
<td>Unidirectional</td>
<td>82.1</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>82.7 ( \pm 0.6 )</td>
</tr>
</tbody>
</table>

Table 7. Study on pseudo-label bank update in phase-2 training.

<table>
<thead>
<tr>
<th>Update</th>
<th>YouTube-VOS 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mathcal{G} )</td>
</tr>
<tr>
<td>✓</td>
<td>82.7 ( \pm 0.5 )</td>
</tr>
</tbody>
</table>

4.4. Discussion

How about more shots? We conduct experiments under the 4-shot and 6-shot settings. We apply our approach to 4- and 6-shot STCN and conduct one round of phase-1 training. Two models achieve the performance of 82.0% and 82.1% on YouTube-VOS 2019, respectively. We further conduct one round of phase-2 training. Both models achieve 82.7% on YouTube-VOS 2019, which is the same as that of 2-shot STCN equipped with our methodology—the performance is already saturated for two-shot VOS and acquiring more labeled data may not be beneficial.

Robustness of our approach. To verify the robustness of our approach, we independently construct five 2-shot VOS datasets from YouTube-VOS 2019 benchmark and train a 2-shot STCN with our methodology on each set. The results are [82.69%, 82.70%, 82.72%, 82.72%, 82.73%], with an average of 82.71% and a standard deviation of 0.015%, showing the robustness of our approach.

5. Conclusion

For the first time, we demonstrate the feasibility that only two labeled frames per video are almost sufficient for training a decent VOS model. On top of this, we present a simple training paradigm to resolve two-shot VOS. The underlying idea behind our approach is to exploit the wealth of information present in unlabeled data in a semi-supervised learning manner. Our approach can be applied to a majority of fully supervised VOS models, such as STCN, RDE-VOS, and XMem. By using 7.3% and 2.9% labeled data of YouTube-VOS and DAVIS benchmarks, our approach achieves comparable results in contrast to the counterparts trained on fully labeled set. With its simplicity and strong performance, we hope our approach can serve as a solid baseline for future research.

References


