

Two-shot Video Object Segmentation

Kun Yan¹ Xiao Li² Fangyun Wei² Jinglu Wang² Chenbin Zhan¹ Ping Wang^{1*} Yan Lu^{2*}

¹Peking University ²Microsoft Research Asia

fkyan2018, zcbin, pwang g@pku.edu.cn fxili11, fawe, jinglwa, yanlu g@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 only 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/yk-pku/Two-shot-Video-Object-Segmentation>.

(a) Previous works on video object segmentation rely on densely annotated videos. We present two-shot video object segmentation, which merely accesses two labeled frames per video.

(b) Comparison among naive 2-shot STCN, STCN trained on full set and 2-shot STCN equipped with our approach on DAVIS 2016/2017 and YouTube-VOS 2018/2019.

Figure 1. (a) Problem formulation. (b) Comparison among STCN variants on various datasets.

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 [7, 9, 21, 30, 37, 46, 52] are trained on densely annotated datasets such as DAVIS [33, 34] and YouTube-VOS [50]. However, acquiring dense annotations, partic-

ularly 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

*Corresponding authors.

satisfactory VOS model on sparsely annotated videos. For it is trained through supervised learning, but each frame the sake of convenience, we use the term 1-shot to denote has either a ground-truth or a pseudo-label attached to it. that N frames are annotated per training video. Note that It is worth noting that, as training progresses, the predic- 1-shot is meaningless since it degrades VOS to the task of image-level segmentation. We use STCN [9] as our base-labels—we update the pseudo-label bank once we identify line due to its simplicity and popularity. Since at least such pseudo labels. The above described process is named two labeled frames per video are required for VOS train- as phase-2 As shown in Fig. 1b, our approach assembled ing, we follow the common practice to optimize a naive onto STCN, achieves comparable results (85.2% v.s 2-shot STCN model on the combination of YouTube-VOS 85.1% on DAVIS 2017, and 82.7% v.s 82.7% on YouTube- and DAVIS, and evaluate on YouTube-VOS 2018/2019 and VOS 2019) in contrast to its counterpart, STCN trained on DAVIS 2016/2017, respectively. We compare the naive full set, though our approach merely accesses 7.3% and shot STCN with its counterpart trained on full set in Fig. 1b. 2.9% labeled data of YouTube-VOS and DAVIS bench- Surprisingly, 2-shot STCN still achieves decent results, for mark, respectively. 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-reference mask of the first frame. Matching-based methods [9, 30, 43, 52] typically employ a memory bank to store label bank for the convenience of access. The VOS model the features of a collection of frames, then a feature matching is then retrained without any restrictions—similar to how

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 85.2%/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 re-tune the networks at test time based on the query mask of the first frame. However, test-time re-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 associated pseudo label that approximates the ground-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 object from the second frame to the last frame. For instance, matching-based methods. STM proposes to construct a STM [30] and STCN [9] take a triplet of frames as input; memory network to store the masks of the previous frames. RDE-VOS [21] and XMem [7] propose to model longer video sequences containing 5 and 8 frames, respectively. 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, 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 ∞ 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 results, these methods need densely annotated videos for phase-1 training since it merely needs a triplet of frames as training. Instead, our method only needs two labeled frames input. Nevertheless, we can train any VOS models in phase-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 “one-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

3.2. Problem formulation and overview

Problem formulation. Given a VOS dataset \mathcal{D} , for each training video $V = [V_1; \dots; V_T]$ 2-D containing T ($T \geq 2$) frames with the associated ground-truth $[Y_1; \dots; Y_T]$, 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 **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

Figure 2. Overview of our methodology. In phase-1 (top), we train a VOS model (STCN) which takes a triplet of frames as input on a two-shot VOS dataset in a semi-supervised manner. We constrain the reference (rst) frame to be a labeled frame to ease the learning. The remaining frames can be either labeled or unlabeled. Then we perform an intermediate inference (middle) to generate pseudo labels for unlabeled frames by the VOS model trained in phase-1, and construct a pseudo-label bank to store the pseudo labels in addition to the ground-truth. In phase-2 (bottom), we re-train a VOS model—which could be most models—on the combination of labeled and pseudo-labeled data without any restrictions on the rst frame. The pseudo-label bank is dynamically updated once more reliable pseudo labels are identified during phase-2 training.

semi-supervised learning, which generates pseudo-labeled where H and W represent the height and the width of the frames together with the labeled ones to enable triplet con- input, $H(\cdot; \cdot)$ denotes the cross-entropy function. Since STCN requires the annotation of the refer- the prediction at pixel $(i; j)$ in the n -th labeled frame, and ence (or rst) frame to segment the object of interest that $Y_n^{(ij)}$ denotes the corresponding ground-truth. appeared in subsequent frames, we always use a labeled The unsupervised loss L_U is a variant of L_S , which is frame as the reference frame to alleviate the error propa- ded as follows:

The last two frames, how- ever, can be either labeled or unlabeled. In our implemen- tation, the last two frames have 0.5 probability of being both unlabeled, and 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 L is the sum of the supervised loss L_S and the unsupervised loss L_U , effected on labeled and unlabeled frames, respectively. L is a standard segmentation loss, which can be formulated as:

$$L_S = \frac{1}{HWN} \sum_{n=1}^N \sum_{i=1}^H \sum_{j=1}^W H(Y_n^{(ij)}; P_n^{(ij)}); \quad (1)$$

$$L_U = \frac{1}{HWN} \sum_{n=1}^N \sum_{i=1}^H \sum_{j=1}^W 1_{[\max(P_n^{(ij)}) > \tau]} H(\hat{Y}_n^{(ij)}; P_n^{(ij)}); \quad (2)$$

where $1_{[\cdot]}$ is the indicator function to filter out the predic- tions whose maximal confidences are lower than the pre- defined threshold τ , $P_n^{(ij)}$ is the prediction at pixel $(i; j)$ in then- n th unlabeled frame, and $\hat{Y}_n^{(ij)} = \text{argmax}(P_n^{(ij)})$ represents the corresponding one-hot pseudo label. By default, we set $\tau = 0.9$ 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.

Figure 3. Illustration of bidirectional inference. Two reference frames are denoted by blue rectangles. A pre-trained VOS model infers unlabeled frames from the inference frame to the end frame and, in a reverse manner, from the inference frame to the beginning frame. We pick the prediction inferred by the labeled frame that is closest to the unlabeled frame.

3.4. Phase-2 training

In phase-1 training, we constrain the reference (or rst) frame to be a labeled frame since the predictions of the subsequent frames significantly rely on the mask of the reference frame. Adopting an unlabeled frame with pseudo-labeled mask as the reference frame aggravates error propagation in the early training. To make full use of the unlabeled data, we present phase-2 training, which lifts the restriction placed on 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.

We perform an intermediate inference before initiating phase-2 training. The inference of a VOS model requires the annotation of the reference (or rst) frame. Nevertheless, 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.

The training process of phase-2 is identical to that of phase-1, except that the reference (or rst) 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 $P^{(ij)}$ to denote the prediction at pixel $(i; j)$. Once the prediction $P^{(ij)}$ meets the condition that $\max(P^{(ij)}) \geq \tau$, where τ denotes a pre-defined threshold, the corresponding pseudo label in pseudo label bank is updated by $y^{(ij)} = \arg\max(P^{(ij)})$. We set $\tau = 0.99$ 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

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 3% and 2.9% labeled data for YouTube-VOS and DAVIS, respectively.

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

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 parameters in random frame

Method	Labeled data	YouTube-VOS 2018					YouTube-VOS 2019				
		G	J _S	F _S	J _U	F _U	G	J _S	F _S	J _U	F _U
STM [30]	100%	79.4	79.7	84.2	72.8	80.9	-	-	-	-	-
MiVOS [8]	100%	80.4	80.0	84.6	74.8	82.4	80.3	79.3	83.7	75.3	82.8
CFBI [52]	100%	81.4	81.1	85.8	75.3	83.4	81.0	80.6	85.1	75.2	83.0
RDE-VOS [21]	100%	-	-	-	-	-	81.9	81.1	85.5	76.2	84.8
HMMN [37]	100%	82.6	82.1	87.0	76.8	84.6	82.5	81.7	86.1	77.3	85.0
JOINT [26]	100%	83.1	81.5	85.9	78.7	86.5	82.7	81.1	85.4	78.2	85.9
STCN [9]	100%	83.0	81.9	86.5	77.9	85.7	82.7	81.1	85.4	78.2	85.9
R50-AOT-L [53]	100%	84.1	83.7	88.5	78.1	86.1	84.1	83.5	88.1	78.4	86.3
XMem [7]	100%	85.7	84.6	89.3	80.2	88.7	85.5	84.3	88.6	80.3	88.6
STCN [9]	100%	83.0	82.0	86.5	77.8	85.8	82.7	81.2	85.4	78.2	86.0
2-shot STCN [9]	7.3%	80.8	79.5	83.9	75.9	84.0	80.6	79.5	83.8	75.6	83.4
2-shot STCN w/ Ours	7.3%	82.9 _{+2 :1}	81.6 _{+2 :1}	86.3 _{+2 :4}	77.7 _{+1 :8}	86.0 _{+2 :0}	82.7 _{+2 :1}	80.9 _{+1 :4}	85.1 _{+1 :3}	78.3 _{+2 :7}	86.6 _{+3 :2}
RDE-VOS [21]	100%	-	-	-	-	-	82.1	81.3	85.7	76.2	85.0
2-shot RDE-VOS [21]	7.3%	-	-	-	-	-	78.4	77.2	81.3	73.4	81.7
2-shot RDE-VOS w/ Ours	7.3%	-	-	-	-	-	82.1 _{+3 :7}	80.4 _{+3 :2}	84.8 _{+3 :5}	77.3 _{+3 :9}	85.8 _{+4 :1}
XMem [7]	100%	85.5	84.4	89.1	80.0	88.3	85.3	84.0	88.2	80.4	88.4
2-shot XMem [7]	7.3%	79.2	77.5	81.9	74.5	82.9	79.1	77.6	81.5	74.5	82.7
2-shot XMem w/ Ours	7.3%	84.8 _{+5 :6}	83.6 _{+6 :1}	88.5 _{+6 :6}	79.2 _{+4 :7}	87.7 _{+4 :8}	84.5 _{+5 :4}	83.5 _{+5 :9}	88.0 _{+6 :5}	79.1 _{+4 :6}	87.3 _{+4 :6}

Table 1. Comparison with different methods on YouTube-VOS 2018 and 2019 validation sets. **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.

Method	Labeled data	DAVIS 2016			DAVIS 2017		
		J & F	J	F	J & F	J	F
STM [30]	100%	89.3	88.7	89.9	81.8	78.2	84.3
CFBI [52]	100%	89.4	88.3	90.5	81.9	79.1	84.6
JOINT [26]	100%	-	-	-	83.5	80.8	86.2
RDE-VOS [21]	100%	91.1	89.7	92.5	84.2	80.8	87.5
MiVOS [8]	100%	91.0	89.6	92.4	84.5	81.7	87.4
HMMN [37]	100%	90.8	89.6	92.0	84.7	81.9	87.5
R50-AOT-L [53]	100%	91.1	90.1	92.1	84.9	82.3	87.5
STCN [9]	100%	91.6	90.8	92.5	85.4	82.2	88.6
XMem [7]	100%	91.5	90.4	92.7	86.2	82.9	89.5
STCN [9]	100%	91.5	90.7	92.3	85.2	81.9	88.5
2-shot STCN [9]	2.9%	87.9	87.1	88.7	81.0	77.7	84.3
2-shot STCN w/ Ours	2.9%	91.3 _{+3 :4}	90.6 _{+3 :5}	92.0 _{+3 :3}	85.1 _{+4 :1}	81.7 _{+4 :0}	88.4 _{+4 :1}
RDE-VOS [21]	100%	91.0	89.5	92.4	84.2	80.7	87.7
2-shot RDE-VOS [21]	2.9%	87.6	86.6	88.8	79.4	75.6	83.1
2-shot RDE-VOS w/ Ours	2.9%	90.8 _{+3 :2}	90.0 _{+3 :4}	92.0 _{+3 :2}	83.9 _{+4 :5}	80.4 _{+4 :8}	87.3 _{+4 :2}
XMem [7]	100%	91.3	90.3	92.4	86.2	82.8	89.7
2-shot XMem [7]	2.9%	88.1	87.1	89.0	81.7	78.2	85.1
2-shot XMem w/ Ours	2.9%	91.3 _{+3 :2}	90.3 _{+3 :2}	92.3 _{+3 :3}	85.6 _{+3 :9}	82.1 _{+3 :9}	89.1 _{+4 :0}

Table 2. Comparisons with different methods on DAVIS 2016 and 2017 validation sets. **-** denotes reproduced result by using the open-source code. By using 2.9% labeled data (2 labeled frames per training video) of DAVIS benchmark, our approach achieves comparable results in contrast to the full-set counterpart, and outperforms the native 2-shot counterpart by large margins.

skipping is gradually increased from 5 to 25 with a curriculum learning schedule. The threshold 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 threshold₂ is set to 0.99.

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;

Components	YouTube-VOS 2019				
	G	J _S	F _S	J _U	F _U
Baseline	80.6	79.5	83.8	75.7	83.4
+phase-1	81.6 :0	79.3	83.5	77.7	86.0
+phase-2	82.7 :1	80.9	85.1	78.3	86.6

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

Pseudo-labeler	YouTube-VOS 2019				
	G	J _S	F _S	J _U	F _U
-	80.6	79.5	83.8	75.7	83.4
STCN	81.2 :6	79.2	83.5	77.2	84.9
MT-STCN	81.6 :4	79.3	83.5	77.7	86.0

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

	YouTube-VOS 2019				
	G	J _S	F _S	J _U	F _U
0.990	81.2	79.4	83.8	76.9	84.5
0.995	81.6	79.3	83.5	77.7	86.0
0.999	81.3	79.4	83.7	76.9	85.2

Table 5. Study of different coef cient used in the MT-STCN.

Figure 4. Study on hyper-parameters α_1 and α_2 , which controls 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 $\alpha_1 = 0.9$ and $\alpha_2 = 0.99$.

(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 con-

ducted 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 hyper-parameters α_1 and α_2 controlling pseudo-labeling in phase-1 and -2, respectively. Fig. 4 displays two accuracy curves by varying α_1 and α_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 $\alpha_1 = 0.9$ and $\alpha_2 = 0.99$ 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: $\theta_t = \alpha \theta_{t-1} + (1 - \alpha) \theta_t$, where θ_t denotes the current iteration, θ_{t-1} and θ_t denote the parameters of MT-STCN and STCN respectively, and α is a weight. It can be seen that using the MT-STCN model surpasses the one without MT strategy. We further ablate 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-

Intermediate inference	YouTube-VOS 2019				
	G	J _S	F _S	J _U	F _U
Unidirectional	82.1	80.8	77.3	77.6	85.2
Bidirectional	82.7 _{+0.6}	80.9	85.1	78.3	86.6

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

Update	YouTube-VOS 2019				
	G	J _S	F _S	J _U	F _U
	82.2	80.7	84.9	77.6	85.5
X	82.7 _{+0.5}	80.9	85.1	78.3	86.6

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

ence to construct a pseudo-label bank to enable phase-2 training. We compare the proposed bidirectional inference with the unidirectional inference, which is typically used in most VOS models. The results are shown in Tab. 6. There is a +0.6% improvement when utilizing bidirectional inference versus unidirectional inference. The reasons are that: (1) some unlabeled frames are not associated with the pseudo labels in the unidirectional inference; (2) the bidirectional inference alleviates the error propagation issue.

Dynamically update the pseudo-label bank. We verify the effectiveness of dynamically updating the pseudo-label bank during phase-2 training, by comparing it with a variant that freezes the pseudo-label bank once constructed. As shown in Tab. 7, freezing the pseudo-label bank slightly hurt the performance. As training progresses, more accurate pseudo labels are generated, thus it is optimal to update the pseudo-label bank to further promote the learning.

Visualization of feature space. We randomly pick two unlabeled frames from the constructed 2-shot YouTube-VOS 2019 training set for feature space visualization. Note we could access their annotations (foreground and background) from the full set. We use PCA to visualize the feature space of naive 2-shot STCN, 2-shot STCN with our training paradigm, and full-set STCN in Fig. 5. Both 2-shot STCN equipped with our methodology, and full-set STCN show more compact clusters.

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 method—the performance is already saturated for two-shot VOS and acquir-

ing 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 82.74% 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

- [1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensemble. *Advances in neural information processing systems*, 27, 2014.
- [2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.
- [3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

- [4] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.
- [5] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2020.
- [6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.
- [7] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv preprint arXiv:2207.07115*, 2022.
- [8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021.
- [9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021.
- [10] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segow: Joint learning for video object segmentation and optical flow. *Proceedings of the IEEE international conference on computer vision*, pages 686–695, 2017.
- [11] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.
- [12] Wenbin Ge, Xiankai Lu, and Jianbing Shen. Video object segmentation using global and instance embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16836–16845, 2021.
- [13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.
- [14] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.
- [15] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021.
- [16] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2019.
- [17] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. *European conference on computer vision*, pages 429–445. Springer, 2020.
- [18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.
- [19] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning*, ICML, volume 3, page 896, 2013.
- [20] Youngjo Lee, Hongje Seong, and Euntai Kim. Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier. *AAAI*, 2022.
- [21] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1332–1341, 2022.
- [22] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 90–105, 2018.
- [23] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *European Conference on Computer Vision*, pages 661–679. Springer, 2020.
- [24] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018.
- [25] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence* 41(6):1515–1530, 2018.
- [26] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9670–9679, 2021.
- [27] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021.
- [28] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 41(8):1979–1993, 2018.
- [29] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7376–7385, 2018.
- [30] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

- [31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [32] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017.
- [33] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.
- [34] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.
- [35] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [36] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. *European Conference on Computer Vision*, pages 629–645. Springer, 2020.
- [37] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021.
- [38] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cscd. *IEEE transactions on pattern analysis and machine intelligence* 38(4):717–729, 2015.
- [39] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
- [40] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.
- [41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.
- [42] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.
- [43] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1296–1305, 2021.
- [44] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017.
- [45] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1140–1148, 2018.
- [46] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021.
- [47] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33:6256–6268, 2020.
- [48] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.
- [49] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.
- [50] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.
- [51] Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, and Stephen Lin. Cross-model pseudo-labeling for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2959–2968, 2022.
- [52] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020.
- [53] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021.
- [54] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7234–7243, 2019.
- [55] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2020.
- [56] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020.

- [57] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.