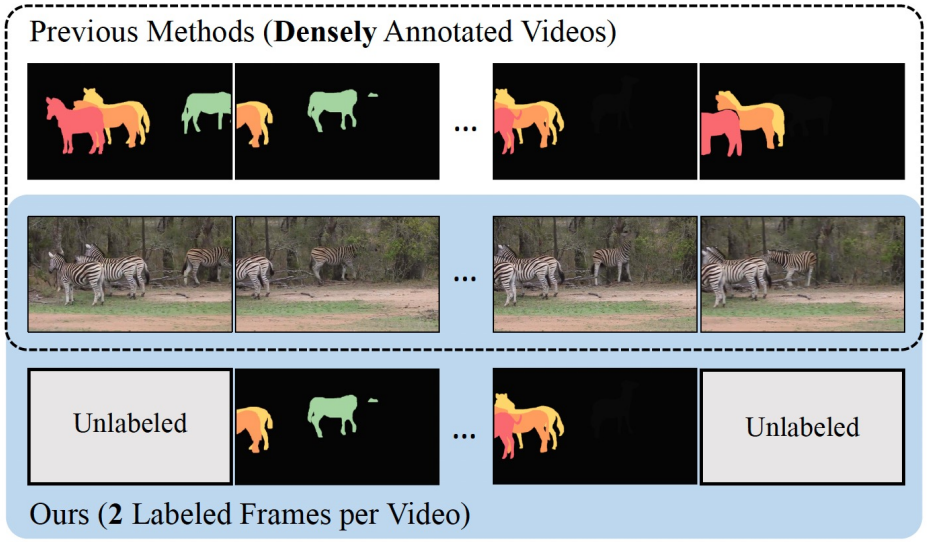
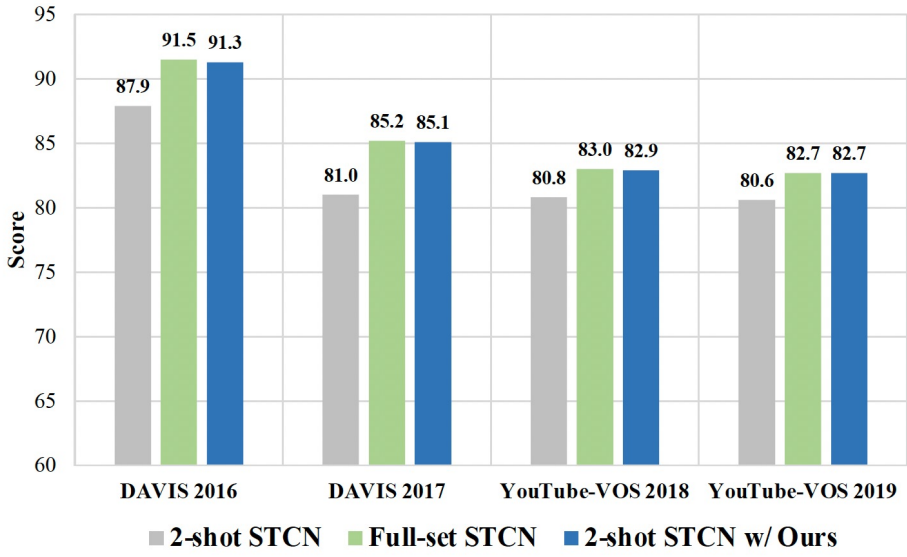
Two-shot Video Object Segmentation

Kun Yan, Xiao Li, Fangyun Wei, Jinglu Wang, Chenbin Zhang, Ping Wang, Yan Lu. “Two-shot Video Object Segmentation”. In CVPR 2023. TUE-AM-216.

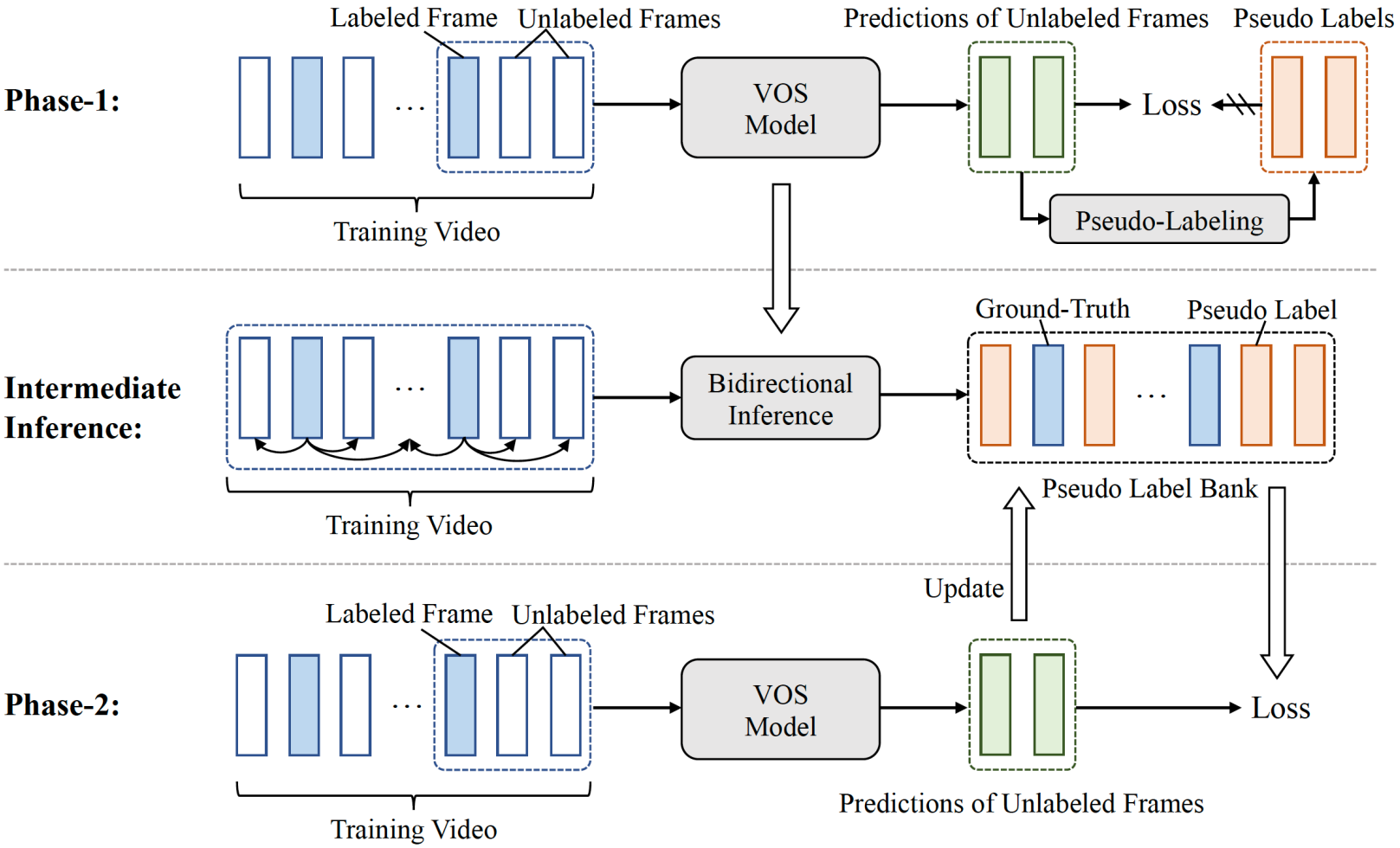


(a) Previous works on video object segmentation rely on densely annotated videos. We present two-shot video object segmentation, which merely accesses two labeled frames per video.



(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.

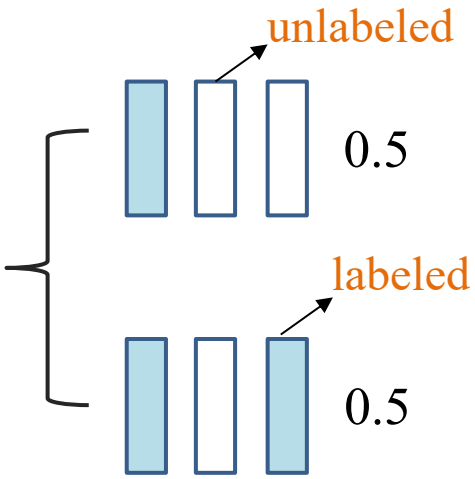
Method



Method

Phase-1:

Sampling:



Training:

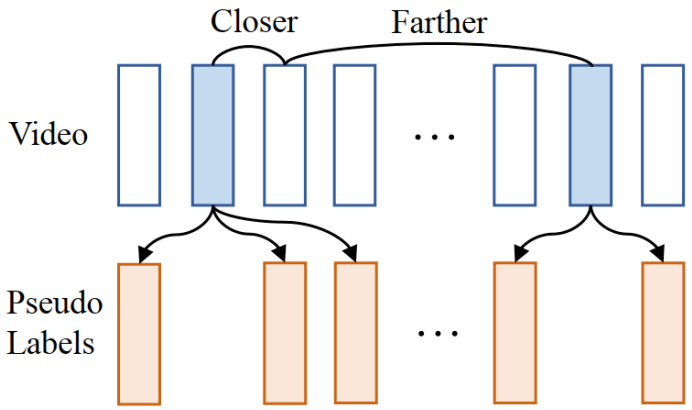
Supervised loss: $$\mathcal{L}_S = \frac{1}{HWN_1} \sum_{n=1}^{N_1} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{H}(\boldsymbol{Y}_n^{(i,j)}, \boldsymbol{P}_n^{(i,j)}),$$

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

Method

Phase-2:

Intermediate inference:



Training:

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

Update pseudo-label bank:

$$\max(\boldsymbol{P}^{(i,j)}) \geq \tau_2 \longrightarrow \hat{\boldsymbol{Y}}^{(i,j)} = \operatorname{argmax}(\boldsymbol{P}^{(i,j)}).$$

Experiments

Main results:

Method	Labeled data	YouTube-VOS 2018					YouTube-VOS 2019				
		$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
STM [28]	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 [50]	100%	81.4	81.1	85.8	75.3	83.4	81.0	80.6	85.1	75.2	83.0
RDE-VOS [20]	100%	-	-	-	-	-	81.9	81.1	85.5	76.2	84.8
HMMN [35]	100%	82.6	82.1	87.0	76.8	84.6	82.5	81.7	86.1	77.3	85.0
JOINT [25]	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 [51]	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* [20]	100%	-	-	-	-	-	82.1	81.3	85.7	76.2	85.0
2-shot RDE-VOS* [20]	**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$_{+4.6}$

Experiments

Ablation study:

Components	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Baseline	80.6	79.5	83.8	75.7	83.4
+phase-1	$81.6_{+1.0}$	79.3	83.5	77.7	86.0
+phase-2	$82.7_{+1.1}$	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.

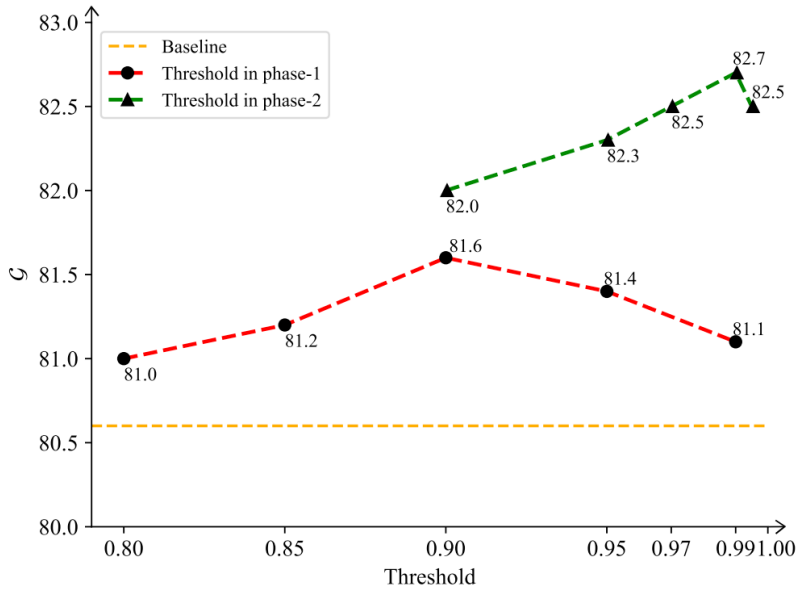


Figure 4. Study on hyper-parameters τ_1 and τ_2, which controls pseudo-labeling in phase-1 and -2, respectively. We adopt a higher threshold in phase-2 training since the predictions in phase-2 are more accurate than that in phase-1. By default, we set $\tau_1 = 0.9$ and $\tau_1 = 0.99$.

Experiments

Ablation study:

Intermediate inference	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
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				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
	82.2	80.7	84.9	77.6	85.5
✓	$82.7_{+0.5}$	80.9	85.1	78.3	86.6

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

Experiments

Discussion:

How about more shots?

Shot	Phase-1	Phase-2
4	82.0	82.7
6	82.1	82.7

Robustness of our approach:

Round	1	2	3	4	5
Phase-2	82.69	82.70	82.72	82.72	82.73

Thanks

北京大學
PEKING UNIVERSITY