上海交通大学
SHANGHAI JIAO TONG UNIVERSITY
浙江大学
ZHEJIANG UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Boosting Video Object Segmentation via Space-time Correspondence Learning

Yurong Zhang[1*], Liulei Li[2*], Wenguan Wang[2†], Rong Xie[1], Li Song[1], Wenjun Zhang[1]

[1]Shanghai Jiao Tong University, [2]Zhejiang University

TUE-AM-215

饮 水 思 源 • 爱 国 荣 校

Preview

Observation

- the weakness of previous matching-based video object segmentation (VOS)
- the potential of self-supervised space-time correspondence learning

Core Idea

- propose a correspondence-aware training framework, which boosts matching-based VOS methods by explicitly encouraging explicit space-time correspondence matching

Performance

- SOTA quantitative outcome
- impressive qualitative results

Contribution

- elegant training framework

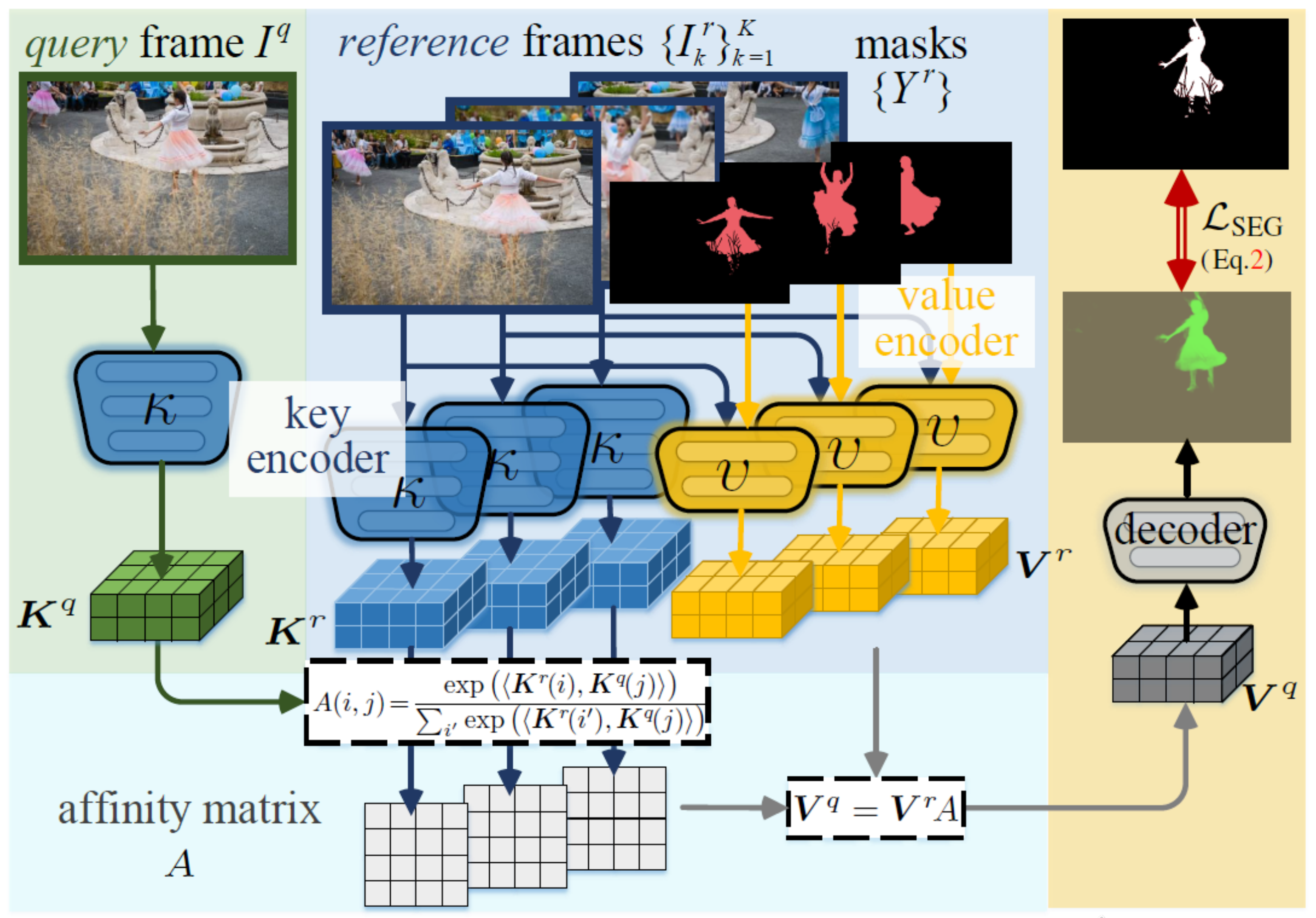
Matching-based VOS Framework

Video Object Segmentation

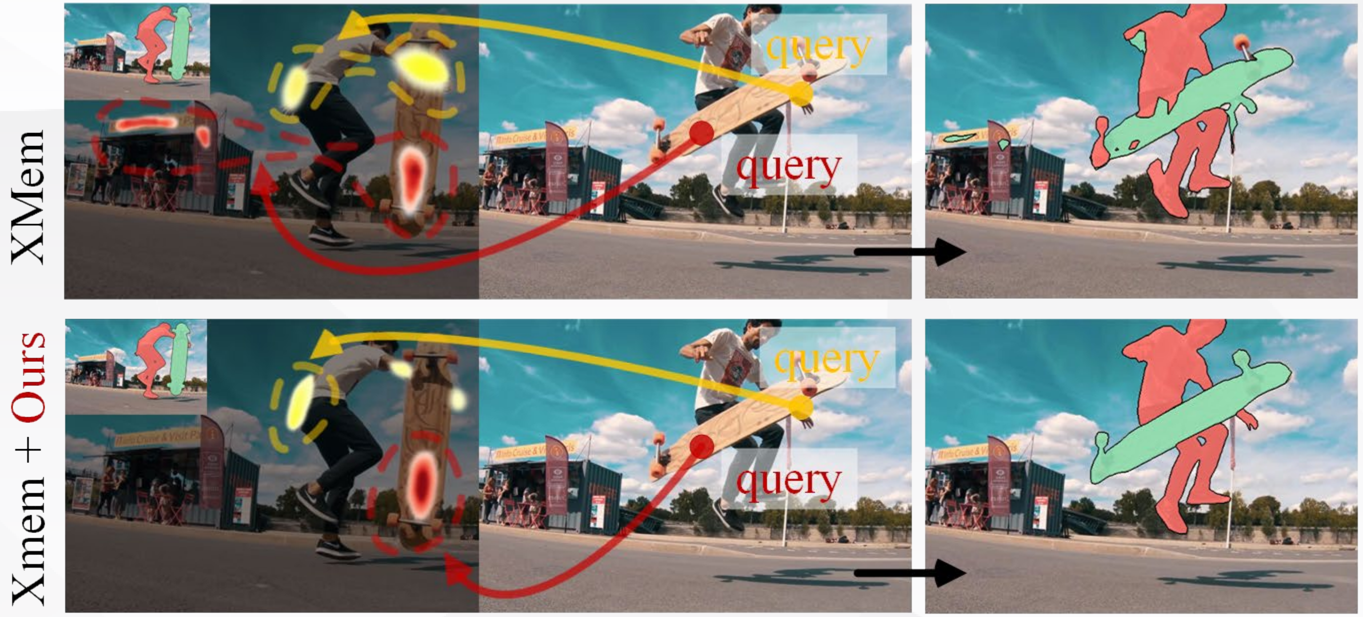
- online-learning based
- propagation-based
- **matching-based**

Matching-based VOS

- explicit object modeling
- current mainstream



Weakness *of previous matching-based VOS (e.g., XMem)*

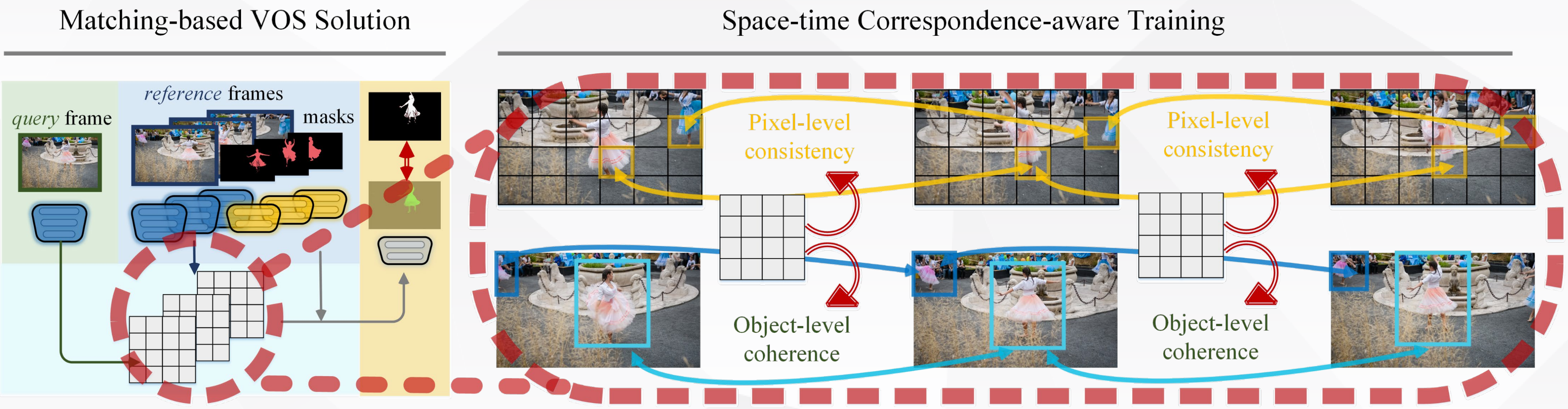


(a) *reference* frame (*w.* mask)
(b) *query* frame
(c) segmentation result of *query* frame

- supervision of gt segmentation masks only
- neglect **explicit constraint** on space-time correspondence learning
- sub-optimal performance (mismatching)

Space-time correspondence-aware training framework

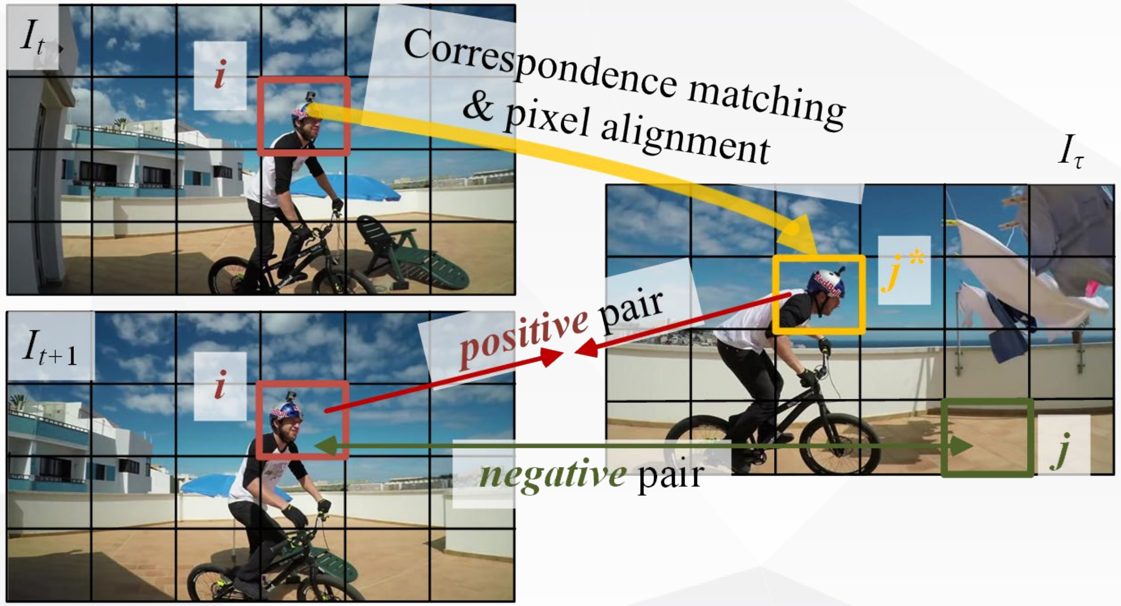
- complementary yet free supervision signals
- pixel-level: spatiotemporally proximate pixels/patches tend to be consistent
- object-level: visual semantics of same object instances at different timesteps tend to retain unchanged.
- deployment friendly

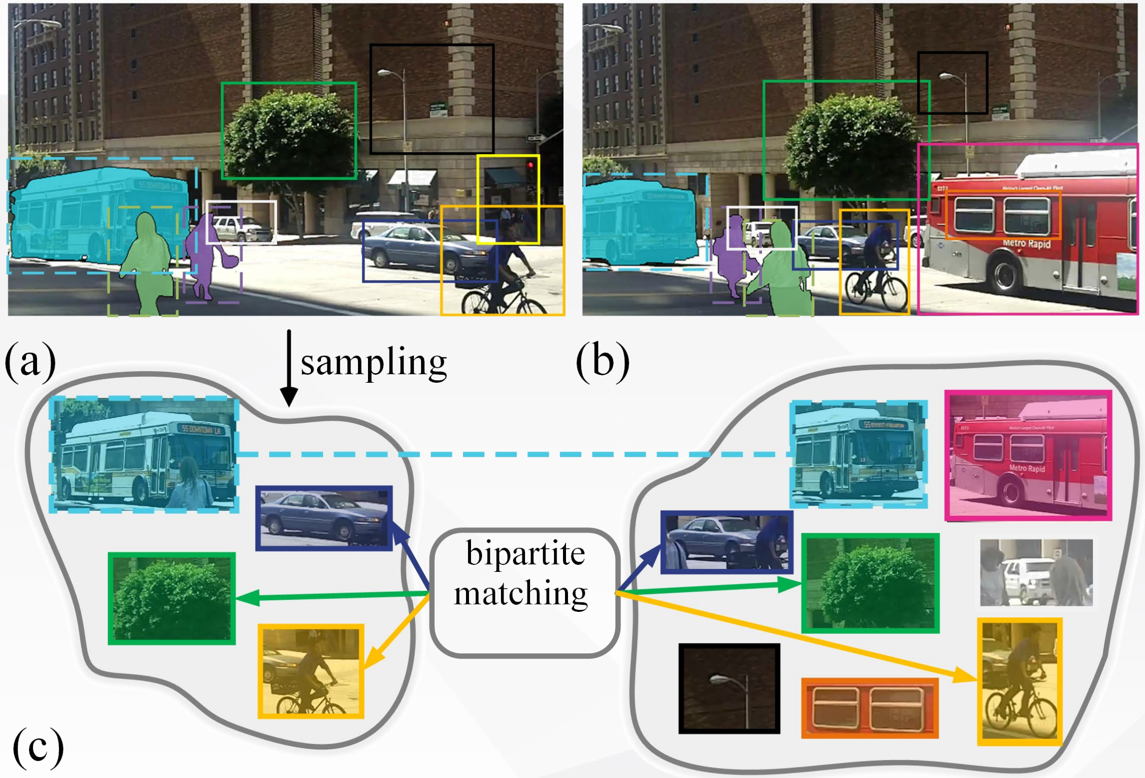


- complement **implicit, segmentation-oriented** supervision signals with **explicit, self-supervised** constraint/regularization over the cross-frame correlation estimation.
- correspondence: pixel-level, object-level

Pixel-level consistency

- local continuity residing in videos
- disambiguate correspondence on both inter- and intra-video levels
- $$\mathcal{L}_{\text{PCL}} = -\log \sum_i \frac{\exp\left(\langle \boldsymbol{K}_{t+1}(i), \boldsymbol{K}_\tau(j^*)\rangle\right)}{\sum_j \exp\left(\langle \boldsymbol{K}_{t+1}(i), \boldsymbol{K}_\tau(j)\rangle\right)}$$





Object-level coherence

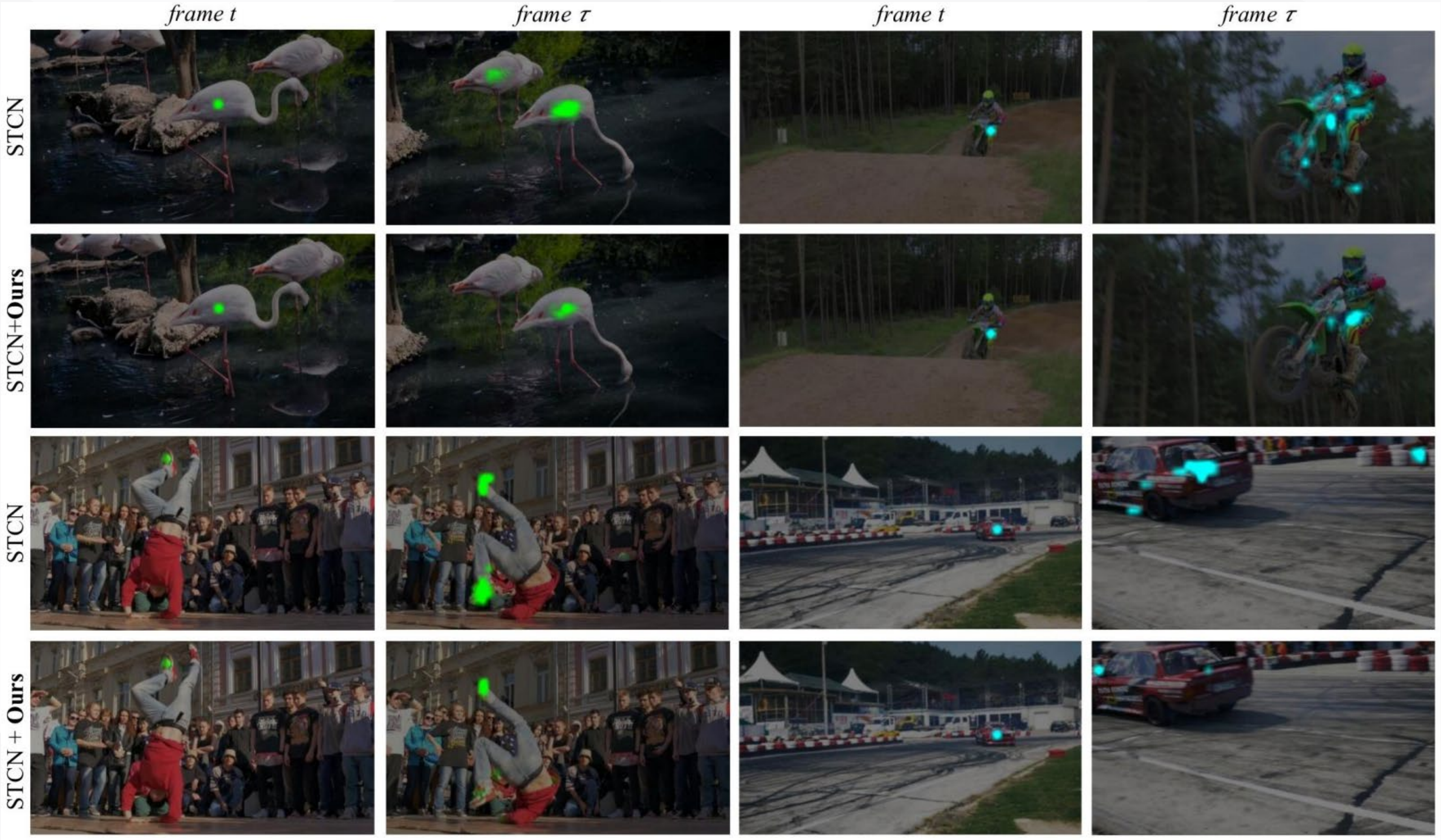
- the content continuity of videos on the object-level
- maximize the similarity of the representations of the same object instance at different timesteps
- $$\boxed{\mathcal{L}_{\mathrm{OCL}} = -\log \sum_{\boldsymbol{p}_i \in \mathcal{Q}} \frac{\exp\left(\langle \boldsymbol{p}_i, \boldsymbol{p}'_{j*} \rangle\right)}{\exp\left(\langle \boldsymbol{p}_i, \boldsymbol{p}'_{j*} \rangle\right) + \sum_{\boldsymbol{o} \in \mathcal{O}} \exp\left(\langle \boldsymbol{p}_i, \boldsymbol{o} \rangle\right)}}$$

Results

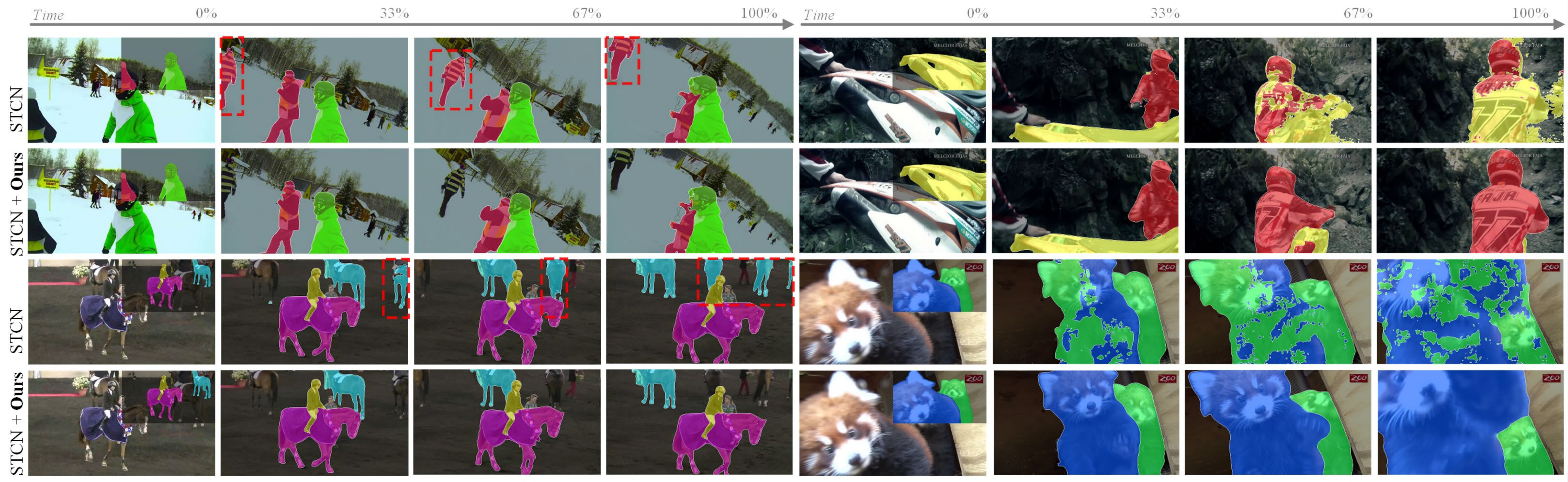
- promote matching-based VOS methods (e.g., STCN and XMem) in a large margin
- further boost SOTA performance

Method	S	DAVIS2017 val	DAVIS2017 test	YouTube-VOS 2018 val	YouTube-VOS 2019 val
SSTVOS [14]	✗	82.5	-	81.7	-
CFBI+ [88]	✗	82.9	75.6	82.8	-
Joint [44]	✗	83.5	-	83.1	-
STCN [12]	✗	82.5	73.9	81.2	-
STCN+Ours	✗	**84.7**	**77.3**	**83.6**	**-**
XMem [10]	✗	84.5	79.8	84.3	-
XMem+Ours	✗	**86.1**	**81.0**	**85.6**	**-**
STM [49]	✓	81.8	72.2	79.4	-
HMMN [58]	✓	84.7	78.6	82.6	82.5
AOT [87]	✓	84.9	79.6	84.1	84.1
PCVOS [51]	✓	86.1	80.2	84.6	84.6
STCN [12]	✓	85.4	76.1	83.0	82.7
STCN+Ours	✓	**86.8**	**79.1**	**85.2**	**84.9**
XMem [10]	✓	86.2	81.0	85.7	85.5
XMem+Ours	✓	**87.7**	**82.2**	**86.9**	**86.8**

Correspondence Matching



Video Object Segmentation Results



Summary

Fresh Insight!

- observe the importance of **explicit supervision signals** for space-time correspondence matching
- take the lead in incorporating **self-constrained correspondence training** target with matching-based VOS

Impressive performance!

- **SOTA** on DAVIS2017 val/test and YouTubeVOS
- improve matching-based VOS in a large margin
- wonderful qualitative result

Charming Framework!

- no modification on network structure
- no extra annotation budget
- no inference time delay and efficiency burden

SJTU

Thanks!

饮 水 思 源　爱 国 荣 校