Deformable VisTR:
Spatio-temporal deformable attention for video instance segmentation

Sudhir Yarram[1], Jialian Wu[1], Pan Ji[2], Yi Xu[2], Junsong Yuan[1]

[1]University at Buffalo [2]OPPO US Research Center, InnoPeak Technology Inc.

22-27 MAY 2022 ICASSP 2022 SINGAPORE

University at Buffalo The State University of New York

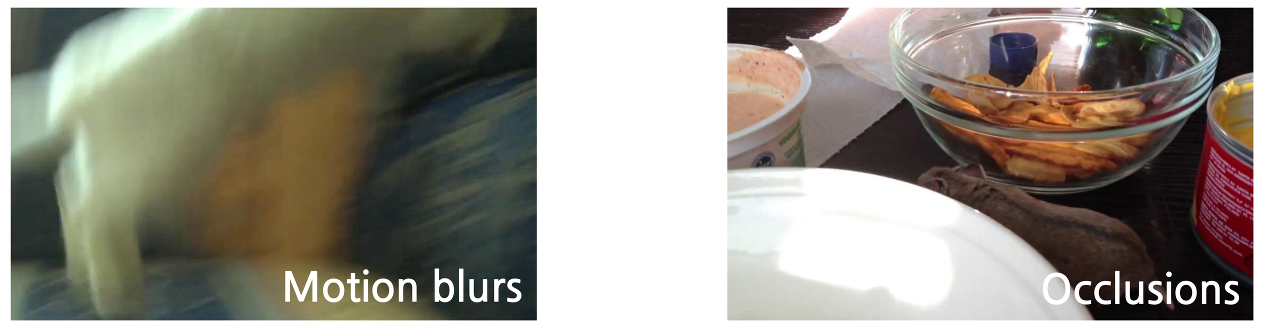
oppo

Github: https://github.com/skrya/DefVIS

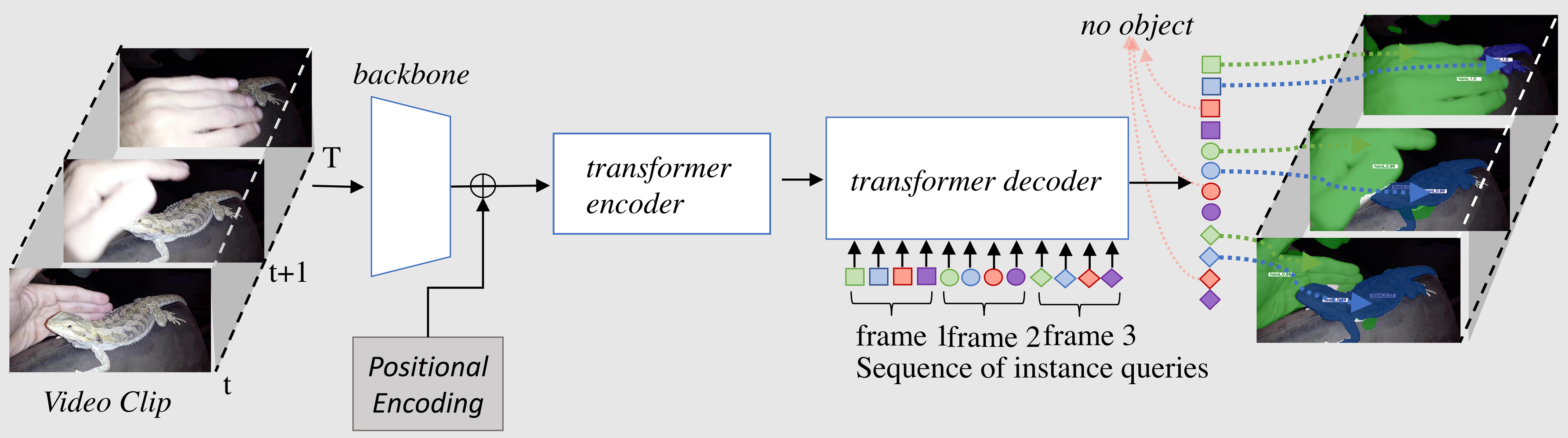
1 Video Instance Segmentation (VIS)

Given a video, predict **spatio-temporal** masks of instances

Typical failure cases in videos

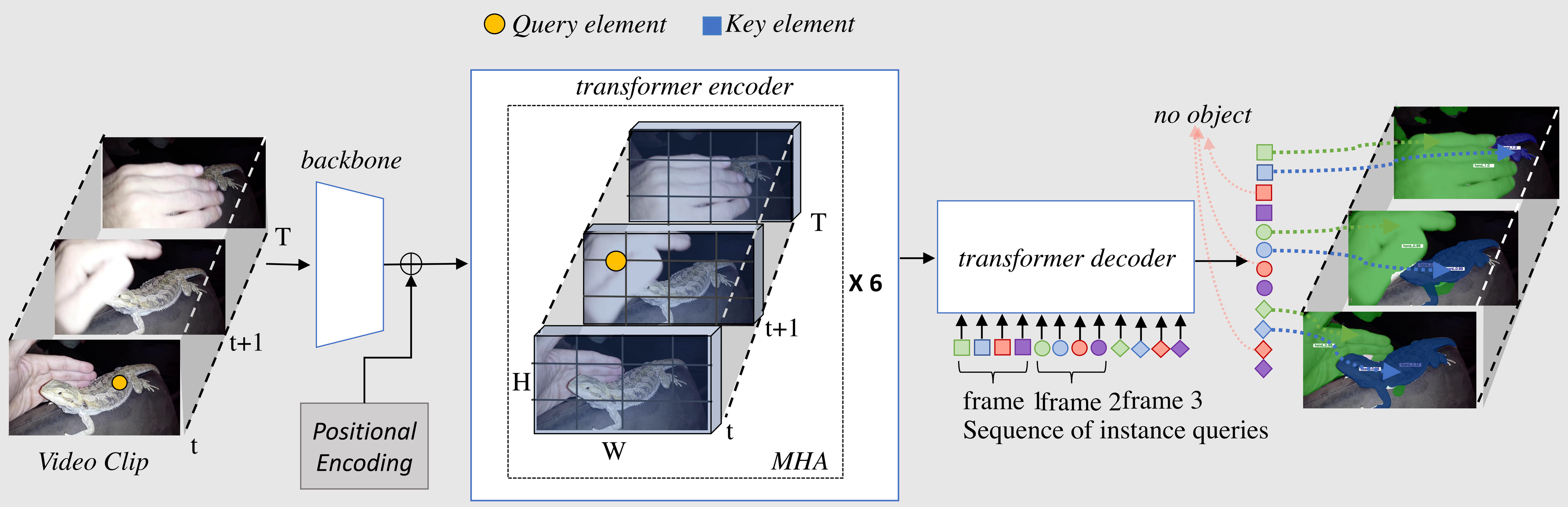


2 VisTR: End-to-End Video Instance Segmentation with Transformers



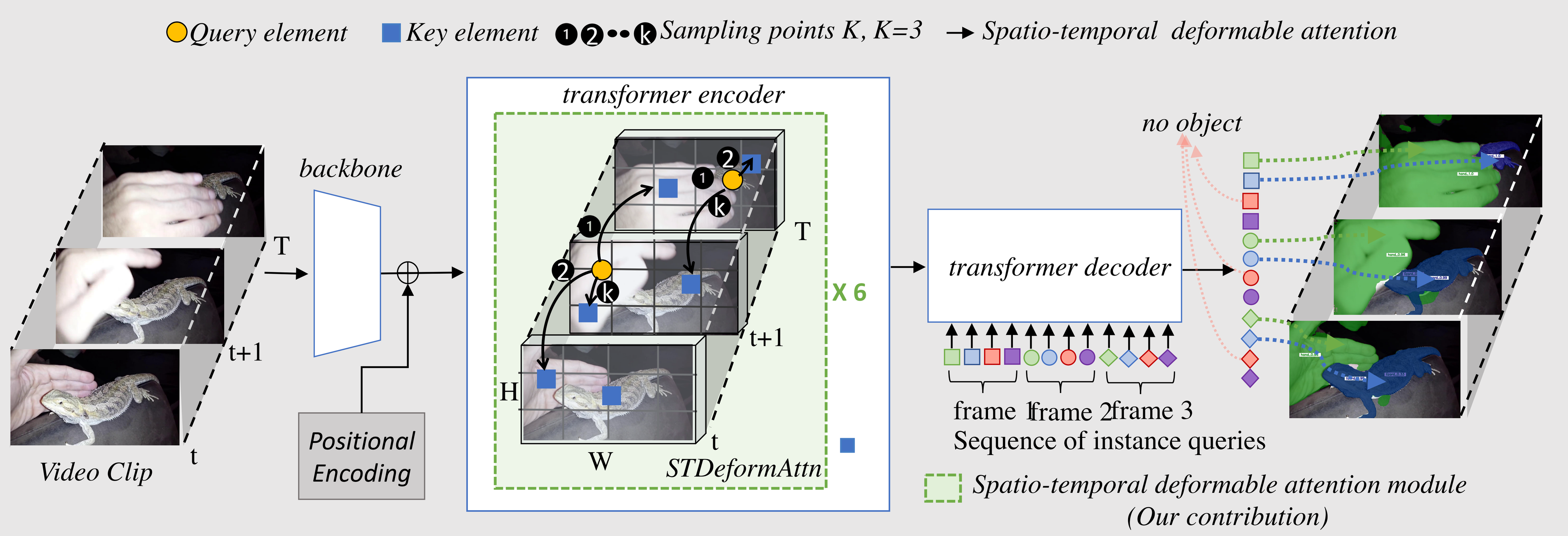
- VisTR takes the entire clip as input and leverages the **transformer** to conduct VIS
- VisTR is slow to converge during training, requiring around **1000 GPU hours** due to the high **computational cost** of its **transformer** attention module.

3 Attention in VisTR



- For each Query element (●), attention is computed with $H \times W \times T$ key elements (■)
- For $H \times W \times T$ query elements the computational complexity for attention is $O(H^2 \times W^2 \times T^2 \times C)$
  - C is the channel dimension of the feature

4 Deformable VisTR



- For each Query element (●), attention is computed with K key elements (■)
- For $H \times W \times T$ query elements the computational complexity for attention is $O(H \times W \times T \times C \times K)$, C is the channel dimension of the feature

Method	Comp. Complexity	Training time (GPU Hours)	Training Epochs	Accuracy (mAP(%))
VisTR [1]	$O(H^2W^2T^2C)$	1000	~500	35.6
Deformable VisTR	$O(HWTCK)$	**120**	**50**	34.6

5 Experiments: Comparison with state-of-the-arts

Method	Fully End-to-End	Aug.	FPS	AP
MaskTrack [2] CVPR'19			28.6	30.3
SipMask [16] ECCV'20		✓	34.1	33.7
STEmSeg [4] ECCV'20		✓✓	4.4	30.6
CompFeat [17] AAAI'21		✓✓	32.8	35.3
SGNet [18] CVPR'21		✓✓	19.8	34.8
STMask [19] CVPR'21			28.6	33.5
CrossVIS [20] ICCV'21			**39.8**	34.8
QueryInst [21] ICCV'21			32.3	34.6
VisTR [1] CVPR'21	✓	✓	30.0	**35.6**
Deformable VisTR	✓	✓	33.0	34.6

- All the entries use ResNet-50 [12] as backbone. The methods are listed in temporal order. " tick" indicates multi-scale input images during training. "double tick" indicates stronger data augmentation (e.g., additional data [17, 4], random crop[3])

6 Experiments: Ablation with different *K*

backbone	K	AP
ResNet-50	16	33.8
ResNet-50	32	34.6

Ablation of STDeformAttn module.

K is the number of key points for each query feature. K = 32 gives the best result.

References:

[1] End-to-End Video Instance Segmentation with Transformers, CVPR'21

[2] Video instance segmentation, CVPR''19

[16] Sipmask: Spatial information preservation for fast im- age and video instance segmentation, ECCV'20

[21] Instances as queries, ICCV'21