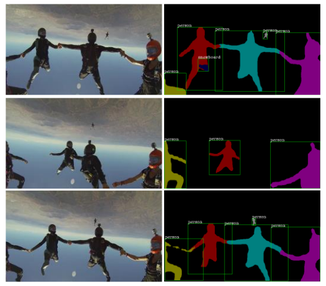
Video Instance Segmentation via Multi-Scale Spatio Temporal Split Attention Transformer

Omkar Thawakar[1] Sanath Narayan[2] Jiale Cao[3] Hisham Cholakkal[1] Rao Muhammad Anwer[1] Muhammad Haris Khan[1]
Salman Khan[1] Michael Felsberg[4] Fahad Shahbaz Khan[1,4]

[1]Mohamed bin Zayed University of AI, UAE [2]Inception Institute of AI, UAE [3]Tianjin University, China [4]Linköping University, Sweden

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022 October 23-27, 2022

Video Instance Segmentaion (VIS)



- VIS is a task of simultaneous classification, segmentation and tracking of object of interest within the videos.
- Youtube-VIS is the large-scale benchmark dataset used for VIS with 2883 high-resolution videos and 40 instance categories.
- Online evaluation is done on validation set with eval metrics *mAP, AP-50, AP-75,AR-1 & AR-10*

Existing Approaches

Video instance segmentation is extended from the traditional image instance segmentation, and aims to simultaneously segment and track all object instances in the video sequence.

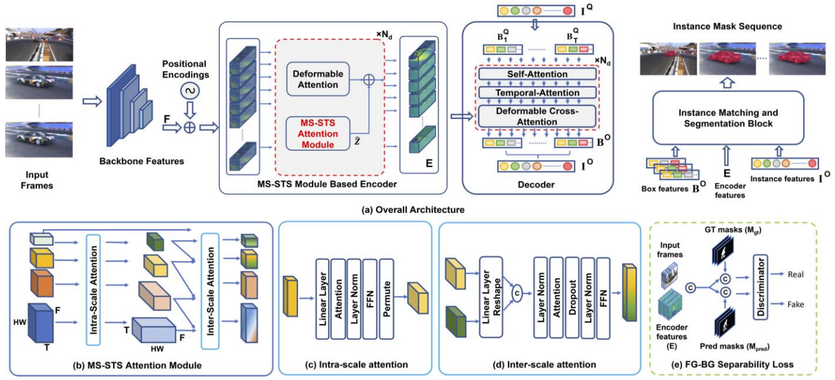
- The baseline method Mask-Track-RCNN [1] is built on top of Mask-RCNN and introduces a tracking head to associate each instance in a video.
- STEM-seg [2] model a video clip as single 3D spatio-temporal and prone to extra computations.
- CrossVIS [3] proposes a learning scheme which used instance features from current frame and localize same instance in other frames.
- VisTR [4], first simple and faster VIS framework based on transformers but using single scale feature hampering the quality of instance segmentation.
- Seqformer [5] near online VIS framework based on Deformable-DETR but not able to tackle instance appearance deformations.

Motivation

- Recent SOTA methods ignores the spatio-temporal feature relationships within multi-scale feature domain during attention computation that is crucial for VIS problem.
- Recent methods strive to predict the accurate instance mask undergoing appearance deformations such as fast motion, scale variations and aspect ratio change in videos.

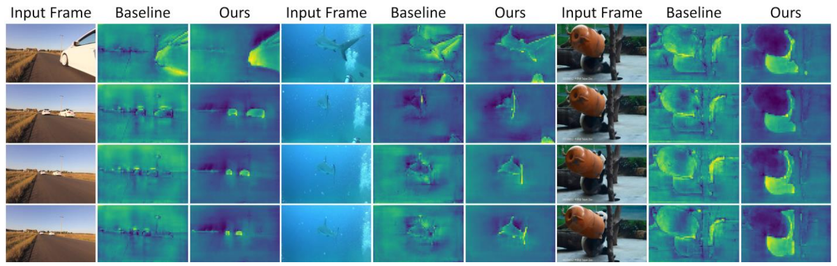
1 Yang *et.al*, Video Instance Segmentation. In ICCV, 2020.
2 Ather *et al*., *Stem-seg*:Spatio temporal embeddinds for instance segmentation in videos. ECCV, 2020.
3 Yang *et al*, *Cross-VIS*: Crossover learning for fast online video instance segmentation. In ICCV, 2021.
4 Wang *et. al. VisTR*: End to end video instance segmentation with transformers. In CVPR, 2021
5 Junfeng *et.al. SeqFormer* : Sequential Transformer for video instance segmentation. In ECCV, 2022.

Proposed MSSTS-VIS model



- ❖ Our MS-STS VIS architecture comprises a backbone, a transformer encoder-decoder and an instance matching and segmentation block. our key contributions are:
- A novel MS-STS attention module in the encoder to capture spatio-temporal feature relationships at multiple scales across frames
- A temporal attention block in the decoder for enhancing the temporal consistency of the box queries.
- A foreground-background (fg-bg) separability loss driven by adversarial training .

Visualization of Attention maps



The baseline struggles to accurately focus on the instance undergoing significant scale variation, where the size becomes extremely small in the later frames. Similarly, it fails to accurately focus on the instance undergoing aspect-ratio change and instance partially visible. In the last video, the baseline inaccurately highlights the irrelevant object (in orange) occluding the target instance. Our MS-STS attention module-based encoder, successfully focuses on these challenging targets despite scale variations (car), aspect-ratio change (shark), partial visibility (person) and appearance deformations due to occlusion (panda).

Experiments

Quantitative analysis

Method	Venue	Backbone	Type	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
IoUTracker+ [26]	ICCV 2019	ResNet-50	-	23.6	39.2	25.5	26.2	30.9
OSMN [27]	CVPR 2018	ResNet-50	Two-Stage	27.5	45.1	29.1	28.6	33.1
DeepSORT [23]	ICIP 2017	ResNet-50	Two-stage	26.1	42.9	26.1	27.8	31.3
FEELVOS [21]	CVPR 2019	ResNet-50	Two-stage	26.9	42.0	29.7	29.9	33.4
SeqTracker [26]	ICCV 2019	ResNet-50	-	27.5	45.7	28.7	29.7	32.5
MaskTrack R-CNN [26]	ICCV 2019	ResNet-50	Two-stage	30.3	51.1	32.6	31.0	35.5
MaskProp [2]	CVPR 2020	ResNet-50	-	40.0	-	42.9	-	-
SipMask-VIS [3]	ECCV 2020	ResNet-50	One-stage	32.5	53.0	33.3	33.5	38.9
SipMask-VIS [3]	ECCV 2020	ResNet-50	One-stage	33.7	54.1	35.8	35.4	40.1
STEm-Seg [1]	ECCV 2020	ResNet-50	-	30.6	50.7	33.5	31.6	37.1
Johnander *et al.*[10]	GCPR 2021	ResNet-50	-	35.3	-	-	-	-
CompFeat [5]	AAAI 2021	ResNet-50	-	35.3	56.0	38.6	33.1	40.3
CrossVIS[28]	ICCV 2021	ResNet-50	One-stage	36.3	56.8	38.9	35.6	40.7
PCAN [11]	NeurIPS 2021	ResNet-50	One-stage	36.1	54.9	39.4	36.3	41.6
VisTR [22]	CVPR 2021	ResNet-50	Transformer	35.6	56.8	37.0	35.2	40.2
SeqFormer [24]	Arxiv 2021	ResNet-50	Transformer	47.4	69.8	51.8	45.5	54.8
MS-STS VIS (Ours)		ResNet-50	Transformer	**50.1**	**73.2**	**56.6**	**46.1**	**57.7**
MaskTrack R-CNN [26]	ICCV 2019	ResNet-101	Two-stage	31.9	53.7	32.3	32.5	37.7
MaskProp [2]	CVPR 2020	ResNet-101	-	42.5	-	45.6	-	-
STEm-Seg [1]	ECCV, 2020	ResNet-101	-	34.6	55.8	37.9	34.4	41.6
CrossVIS [28]	ICCV 2021	ResNet-101	One-stage	36.6	57.3	39.7	36.0	42.0
PCAN [11]	NeurIPS 2021	ResNet-101	One-stage	37.6	57.2	41.3	37.2	43.9
VisTR [22]	CVPR 2021	ResNet-101	Transformer	38.6	61.3	42.3	37.6	44.2
SeqFormer [24]	Arxiv 2021	ResNet-101	Transformer	49.0	71.1	55.7	46.8	56.9
MS-STS VIS (Ours)		ResNet-101	Transformer	**51.1**	**73.2**	**59.0**	**48.3**	**58.7**
SeqFormer [24]	Arxiv 2021	Swin-L	Transformer	59.3	82.1	66.6	51.7	64.4
MS-STS VIS (Ours)		Swin-L	Transformer	**61.0**	**85.2**	**68.6**	**54.7**	**66.4**

State-of-the-art comparison on YouTube-VIS 2019 val set. Our MRSTS VIS consistently outperforms the state-of-the-art results reported in literature.

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN[25]	28.6	48.9	29.6	26.5	33.8
SipMask-VIS[3]	31.7	52.5	34.0	30.8	37.8
VisTR[22]	31.8	51.7	34.5	29.7	36.9
CrossVIS[28]	34.2	54.4	37.9	30.4	38.2
IFC[8]	36.6	57.9	39.9	–	–
SeqFormer[24]	40.5	62.4	43.7	36.1	48.1
MS-STS VIS (Ours)	**42.2**	**63.7**	**46.5**	**41.7**	**51.1**

State-of-the-art comparison on YouTube-VIS 2021 val set. All results are reported using the same ResNet-50 backbone. Our MS-STS achieves state-of-the-art results with an overall mask AP of 42.2% and an absolute gain of 2.8% over the best existing SeqFormer at a higher overlap threshold of AP75.

Qualitative results

Conclusion

We proposed MS-STS VIS, which comprises a novel multi-scale spatio-temporal split attention (MS-STS) module to effectively capture spatio-temporal feature relationships at multiple scales across frames in a video. We further introduced an auxiliary discriminator network during training that strives to enhance fg-bg separability within the multi-scale spatio-temporal feature space. Our MS-STS VIS specifically tackles target appearance deformations due to real-world challenges such as, scale variation, aspect-ratio change and fast motion in videos.