Video Instance Segmentation via Multi-scale Spatio-temporal Split Attention Transformer (Supplementary Material)

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]MBZUAI, UAE [2]IIAI, UAE
[3]Tianjin University, China [4]Linköping University, Sweden

In this supplementary, we present additional quantitative and qualitative results to further validate the efficacy of our proposed multi-scale spatio-temporal split attention based video instance segmentation (MS-STS VIS) framework. The additional ablation studies w.r.t. different design choices are presented in Sec. S1 followed by additional qualitative results in Sec. S2. We discuss how different settings impact the proposed MS-STS VIS performance. All the experiments are conducted on Youtube-VIS [5] dataset. For fair evaluation, we follow the same settings of [3] for baseline and our model. We choose ResNet-50[1] as our backbone for all the ablation experiments shown. Additional qualitative results is reported on Youtube-VIS [5,4].

S1 Additional Ablation Analysis

S1.1 Encoding Spatio-temporal Attention

Design of attention module: We note that baseline [3] only employs spatial attention (SA) across scales and ignores temporal attention (TA). **(i)** While a *joint* multi-scale spatio-temporal attention can address this, it is memory-wise prohibitive. **(ii)** A memory-efficient alternative is a *dis-joint* attention, where TA is computed separately (our *intra-scale*) and then combined with SA. While this improves the results, it does not explicitly capture joint spatio-temporal (ST) information likely to aid in VIS task. This motivates us to introduce a light-weight ST attention (*inter-scale*) that efficiently attends to all scales, taking two successive scales at a time. Our hybrid MS-STS with both intra- and inter-scale attention achieves significant gain of 2.0% over the baseline. Furthermore, using ST attention at inter-scale, as in our MS-STS, yields better results compared to using it at intra-scale. Consequently, since our carefully designed inter-scale attention attends to only two scales at a time and our intra-scale temporal attention causes negligible overhead (fewer tokens to attend), we observe only a minor drop in inference speed compared to the baseline (**Ours:** 10 FPS *vs.* Baseline [3]: 11 FPS).

Encoder Variants Integrating Spatio-temporal Attention: Next, we ablate the encoder design variants integrating spatio-temporal attention on the

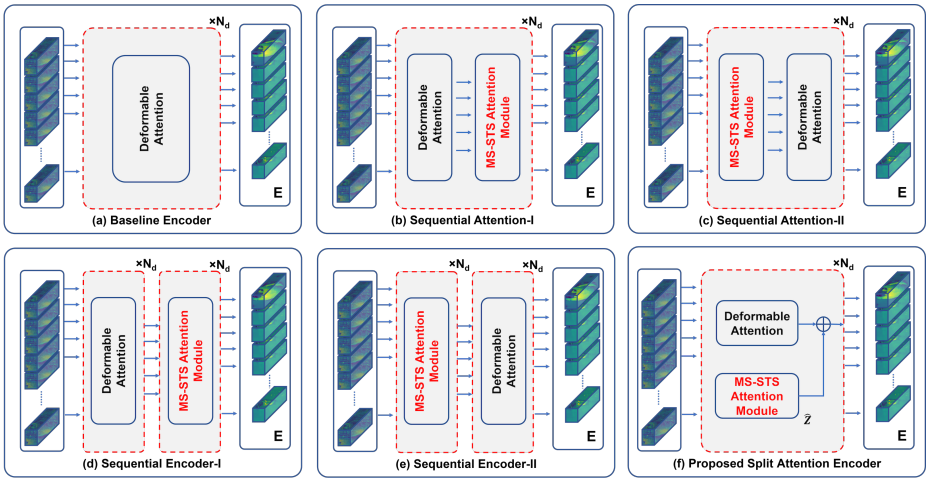

Fig. S1: Encoder variants integrating spatio-temporal attention. (a) The baseline encoder with standard deformable attention. In (b) Sequential Attention-I and (c) Sequential Attention-II, the MS-STS attention is integrated into the standard deformable attention (a) sequentially in each encoder layer. Similarly, in Sequential Encoder-I (d) and Sequential Encoder-II (e), the proposed MS-STS encoder (N_d attention layers) is placed sequentially after or before the standard deformable encoder. Finally, in (f), we show the proposed Split Attention Encoder, where our MS-STS attention is in parallel to the deformable attention in every encoder layer.

Youtube-VIS 2019 [5] val. set. For this ablation, we only consider the variations w.r.t. the encoder and exclude our other contributions (*i.e.*, temporal consistency in decoder and foreground-background (fg-bg) separability) from our final VIS framework. Fig. S1 presents the encoder design variations integrating spatio-temporal attention. The baseline encoder with the standard multi-scale deformable spatial attention is shown in Fig. S1(a). We integrate our proposed MS-STS attention module in a sequential manner after and before the baseline deformable attention in each layer of the encoder and refer to these variants as Sequential Attention-I (Fig. S1(b)) and Sequential Attention-II (Fig. S1(c)). Similarly, we refer to the variants with sequentially placed encoders (N_d layers together form an encoder) as Sequential Encoder-I (Fig. S1(d)) and Sequential Encoder-II (Fig. S1(e)). Finally, our proposed split attention based encoder, where our MS-STS attention module is in parallel to the standard deformable attention in each encoder layer is shown in Fig. S1(f). The VIS performance of each of the variants is presented in Tab. S1. The proposed split attention encoder achieves the best performance with an absolute gain of 2.0% in terms of overall mask AP, over the baseline encoder.

Table S1: VIS performance comparison on Youtube-VIS 2019 val. set, with encoder variants integrating spatio-temporal attention. All results are reported using the same ResNet-50 backbone. Note that here we only analyze the encoder variants and exclude our other contributions (*i.e.*, temporal consistency in decoder and foreground-background (fg-bg) separability). Our proposed split attention encoder achieves the best performance over the other variants considered, since it effectively encodes the multi-scale spatio-temporal feature relationships that are crucial to tackle target appearance deformations in videos. Notably, the proposed split attention encoder achieves a significant gain in performance at a higher overlap threshold of AP_{75}. See Sec. S1.1 and Fig. S1 for more details.

Method	**AP**	**AP_{50}**	**AP_{75}**	**AR_1**	**AR_{10}**
(a) Baseline Encoder	46.4	68.7	50.3	44.9	54.3
(b) Sequential Attention-I	47.0	69.1	51.5	45.4	54.8
(c) Sequential Attention-II	47.3	69.3	51.8	45.6	55.1
(d) Sequential Encoder-I	46.8	68.8	51.3	45.3	54.6
(e) Sequential Encoder-II	47.1	68.9	51.4	45.5	54.4
(f) Proposed Split Attention Encoder	**48.4**	**70.4**	**54.8**	**45.9**	**56.1**

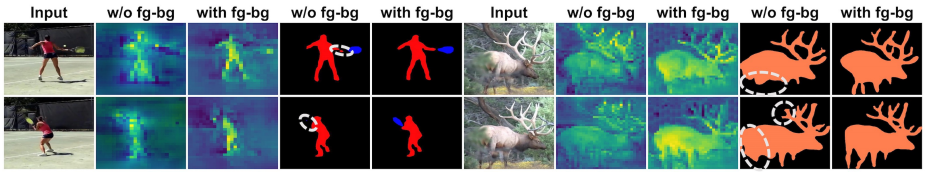

Fig. S2: Example encoder attention map along with final mask w/o and with fg-bg loss. Our loss aids fg-bg separability and mask plausibility for confusing (left: behind *person* & *racket*) and cluttered (right: tree branches around *deer*) backgrounds, leading to better masks compared to w/o fg-bg loss (white dotted regions).

S1.2 Foreground-Background Separability Loss

To complement overall goal of class-specific segmentation, our fg-bg loss aims to enhance separation between fg-bg regions in encoder feature space in a *class-agnostic* manner. The enhanced encoder features aid in improved decoding of target instances. We conjecture that in the aforementioned class-agnostic (binary) setting, both fg-bg separability and mask plausibility (Fig. S2) strive for a common objective leading to an absolute gain of 1.0% on the final VIS task (Tab. 3:left in the main paper).

S1.3 Aggregation of Temporal Information

Our proposed MS-STS attention module explicitly correlates and aggregates temporal information within multiple frames to learn video level instance fea-

Table S2: Effect of input frames on baseline *vs.* proposed MS-STS VIS framework.

Method	Input Frames	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Baseline	2	38.4	58.7	40.1	36.6	42.1
MS-STS (ours)	2	**40.7**	**63.3**	**43.7**	**41.0**	**49.6**
Baseline	3	41.7	64.7	45.5	42.6	50.3
MS-STS (ours)	3	**44.4**	**67.5**	**48.6**	**45.4**	**53.1**
Baseline	4	44.3	69.1	49.4	44.2	52.3
MS-STS (ours)	4	**47.5**	**71.5**	**51.7**	**45.7**	**56.1**
Baseline	5	46.4	68.7	50.3	44.9	54.3
MS-STS (ours)	5	**50.1**	**73.2**	**56.6**	**46.1**	**57.7**

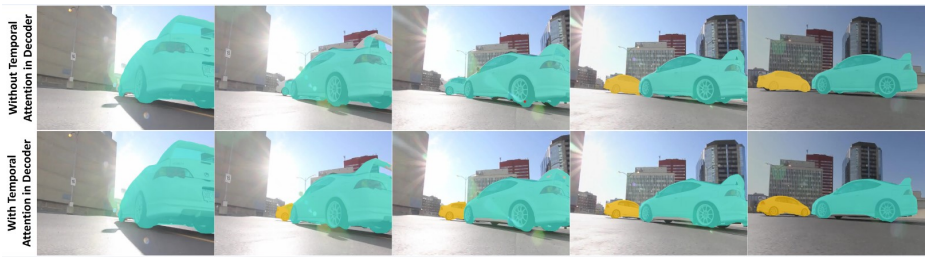

Fig. S3: Our temporal attention in decoder improves mask consistency, particularly, second *car* (in yellow) in frames 2 and 3.

tures. In Tab. S2, we analyse the effect of using fewer frames for instance segmentation. As presented in Tab. S2, our proposed MS-STS attention shows significant improvement with increase number of input frames over the baseline.

Furthermore, our temporal attention in decoder aims to improve the temporal consistency of queries. It is computed as standard self-attention on box-queries of an instance across T frames. Here, Fig. S3 shows that the temporal mask consistency improves with temporal attention in decoder.

S1.4 A Note on the Video Length *T*:

Following baseline SeqFormer [3], we use T=5. Since videos are annotated at 6 fps (every 5^{th} frame in 30fps video) in Youtube-VIS, T=5 $\approx$ 0.83 sec, during which we observe target appearance deformations to occur. Further, our approach maintains gain of >2.0% over baseline, when using a higher T=36 with a reduced spatial resolution, as in VisTR [2].

S2 Additional Qualitative Results

Our method achieves favorable performance by accurately associating and segmenting object instances under fast motion, *e.g.*, rows 2, 3, 6, 7 in Fig. S4 and

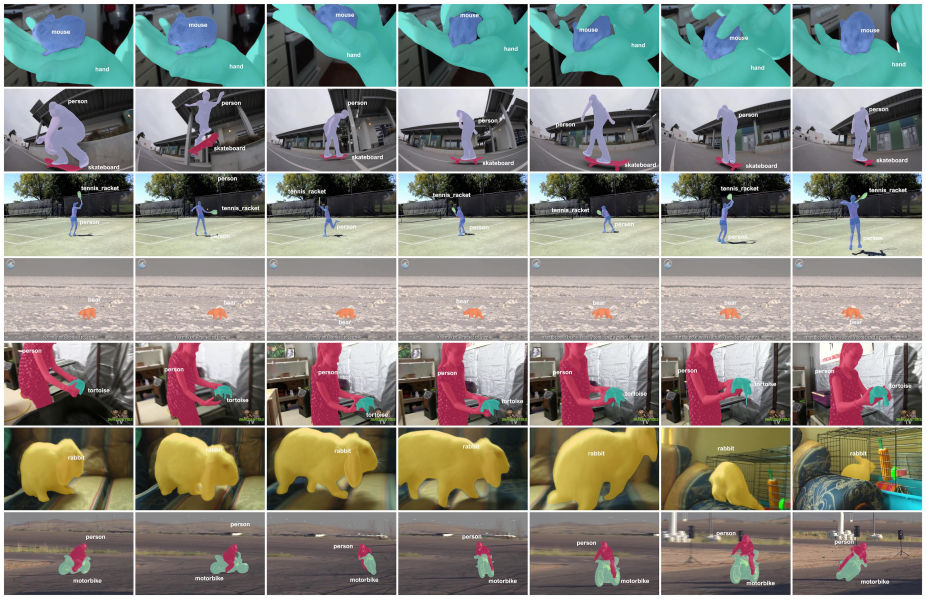

Fig. S4: Additional Qualitative results obtained by our MS-STS VIS framework on seven example videos in the Youtube-VIS 2019 val set. Our MS-STS VIS achieves promising video mask prediction in various challenging scenarios including, fast motion (*person, skateboard* in row 2, *rabbit* in row 6, *person, motorbike* in row 7), scale change (*rabbit* in row 6, *person, motorbike* in row 7), aspect-ratio change (*mouse* in row 1, *person* in rows 2, 3). Also see the videos in https://github.com/OmkarThawakar/MSSTS-VIS.

rows 2 to 4 in Fig. S5. Notably, we can observe that our approach successfully tracks and segments the true object instance and not its shadow/reflection Fig. S4 (row 6) and Fig. S5 (row 6). Furthermore, Fig. S4 (rows 1, 2, 6, 7) and Fig. S5 (rows 1 to 7) show the performance of our proposed approach, when the target object instances undergo changes in aspect ratio and size. Our method reliably tracks and segments the object instances despite these changes in aspect ratio and size. Fig. S4 (rows 1, 5) display qualitative results under occlusion. Our proposed method accurately segments and tracks objects, such as *mouse*, *tortoise* and *rabbit* in these examples. Furthermore, Fig. S6 shows a qualitative comparison between our MS-STS VIS and other recent approaches on two example videos from Youtube-VIS 2019 val. set.

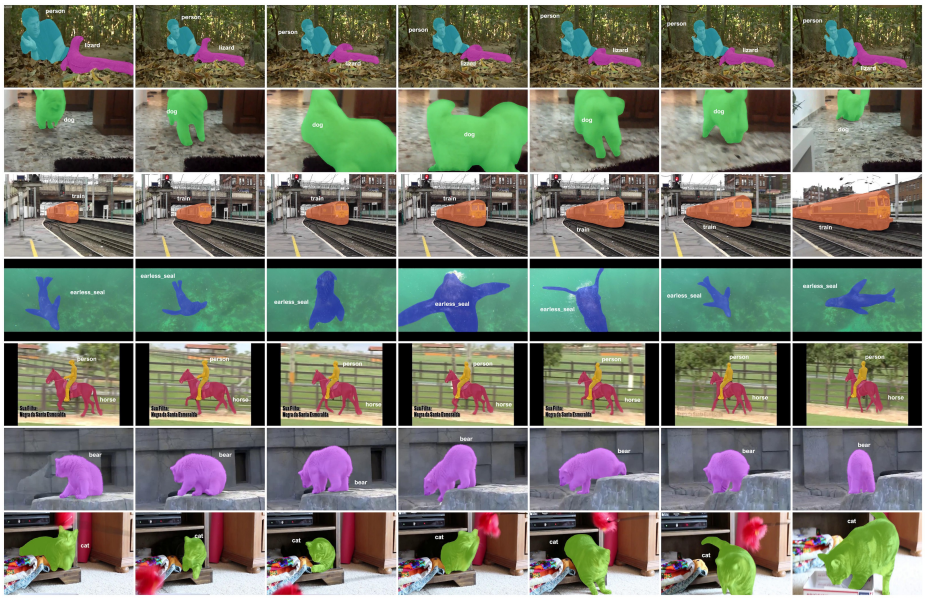

Fig. S5: Qualitative results on seven example videos in the Youtube-VIS 2021 val set. Our MS-STS VIS achieves favorable video mask prediction in various scenarios involving target appearance deformations: fast motion (*dog* in row 2, *train* in row 3, *earless seal* in row 4), scale variation (*person* in row 1 and *cat* in row 7), aspect-ratio change (*bear* in row 6). Also see the videos in https://github.com/OmkarThawakar/MSSTS-VIS.

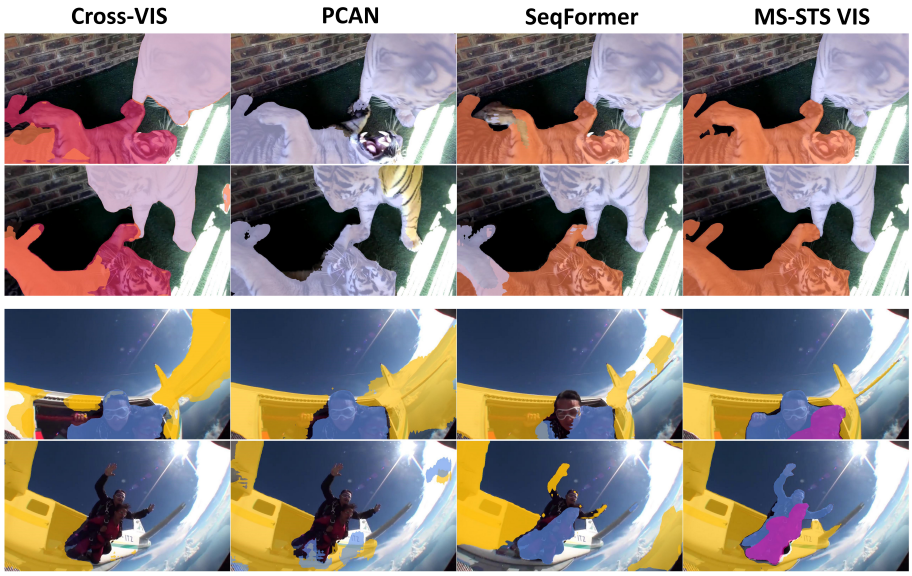

Fig. S6: Comparing MS-STS VIS with other recent methods.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 1
2. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR (2021) 4
3. Wu, J., Jiang, Y., Zhang, W., Bai, X., Bai, S.: Seqformer: a frustratingly simple model for video instance segmentation. In: arXiv preprint arXiv:2112.08275 (2021) 1, 4
4. Xu, N., Yang, L., Yang, J., Yue, D., Fan, Y., Liang, Y., Huang, T.S.: Youtube-vis dataset 2021 version. https://youtube-vos.org/dataset/vis (2021) 1
5. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019) 1, 2