TDViT: Temporal Dilated Video Transformer for Dense Video Tasks

Guanxiong Sun[1,2], Yang Hua[1], Guosheng Hu[2], and Neil Robertson[1]

[1] EEECS/ECIT, Queen's University Belfast, UK
[2] Oosto, Belfast, UK
{gsun02,y.hua,n.robertson}@qub.ac.uk, huguosheng100@gmail.com

1 Speed-accuracy Trade-off of TDViT

We compare TDViT with various widely used backbones, including R-50/101, Swin-T/S/B, on video object detection. The results are tested with the single-frame FasterRCNN [3] on a Tesla V100 GPU. The metric of average precision AP^{box} is evaluated on the ImageNet VID [4] validation set. All TDViT variants greatly surpass other architectures with comparable model sizes, i.e., number of parameters denoted as #Param. For *basic* variants, TDViT-T achieves 49.1% of AP^{box}, outperforming R-50/Swin-T (44.3/47.1%) by +4.8/2.0% of AP^{box}, respectively. TDViT-S/B achieves 55.4/56.0% of AP^{box}, outperforming Swin-S/B (52.6/53.2%) by +2.8% and TDViT-S outperforms R-101 (48.5%) by +6.9% of AP^{box}. We also introduce three *advanced* variants of TDViT, called TDViT-T^+/S^+/B^+, by adding two extra TDTBs in the stage3 of their corresponding basic variants. As a result, the performance is improved to 50.9/55.7/56.1% of AP^{box}_{50}, respectively. As shown in Figure S1, TDViT achieves better speed-accuracy trade-offs.

2 Detailed Spatiotemporal Scheme on Different Variants

The detailed specifications of spatiotemporal schemes for different variants are shown in Table S1. We use the split spatiotemporal scheme by default because it achieves better performance than the factorised scheme. For better understanding, we also list the architecture of Swin [2], whose basic building block is space-only transformers. The space-only transformers and the proposed temporal dilated transformers (TDTB) are denoted as s and t, respectively. The number of parameters is denoted as #Param.

3 Qualitative Results

We also conduct qualitative experiments to better explain the mechanism of our TDViT on two different benchmarks, the ImageNet VID [4] for video object detection and the YouTube VIS [5] for video instance segmentation. The example frames are from validation sets. The visualisation results are obtained

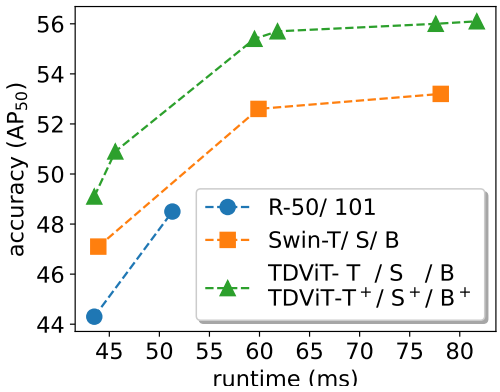

Fig. S1. Speed-accuracy trade-offs between TDViT and other widely used backbones. Compared to ResNet [1] and Swin [2], our approach achieves better performance with similar complexity.

Table S1. The detailed architecture specifications of different variants using the split spatiotemporal scheme.

Variants	stage 1		stage 2		stage 3		stage 4		#Param
	s	t	s	t	s	t	s	t	
Swin-T	2	0	2	0	6	0	2	0	47.8 M
Swin-S/B	2	0	2	0	18	0	2	0	69.1/107.1 M
TDViT-T	1	1	1	1	3	3	1	1	47.8 M
TDViT-S/B	1	1	1	1	9	9	1	1	69.1/107.1 M
TDViT-T$^+$	1	1	1	1	3	5	1	1	51.3 M
TDViT-S$^+$/B$^+$	1	1	1	1	9	11	1	1	72.7/113.5 M

with FasterRCNN [3] and MaskTrack RCNN [5], respectively. We compare our TDViT-T with its 2D couterpart Swin-T[2] to demonstrate the effectiveness of our method.

3.1 Video Object Detection

Figure S2 shows some typical examples where our TDViT-T improves object detection results compared to the baseline Swin-T. The columns (a) and (b) show the detected bounding boxes of Swin-T and TDViT-T, respectively. There are many typical cases where the still image backbones are struggle. For example, when the turtle is partially occluded, the detection scores are drastically decreased (low confidence) and the bounding boxes are not accurate neither. In some blurred frames, the 2D backbones classifies the blurred turtle into a wrong category (false positive) and sometimes recognise it as background (false

negative). In the contrast, our TDViT can model long-range spatiotemporal information, and therefore it can utilise temporal information from the memory to enhance the aforementioned bad frames. As shown in Figure S2 (b), TDViT-T can generate accurate bounding boxes in these typical scenarios, which demonstrate the effectiveness of our TDViT.

3.2 Video Instance Segmentation

Figure S3 shows some typical examples where our TDViT-T improves instance segmentation results compared to the baseline Swin-T. The columns (a) and (b) show the segmentation masks and bounding boxes of Swin-T and TDViT-T, respectively. Similar to video object detection, our TDViT works better than its 2D counterpart, Swin, in occluded and blurred scenarios. TDViT-T generates more accurate results than Swin-T, which further demonstrates the effectiveness and compatibility of our TDViT.

Fig. S2. Typical cases for video object detection. (a) and (b) show the detection results of Swin-T and TDViT-T, respectively. In occlusion and blur scenarios, the performance of Swin-T is drastically decreased, while our TDViT-T can still work well.

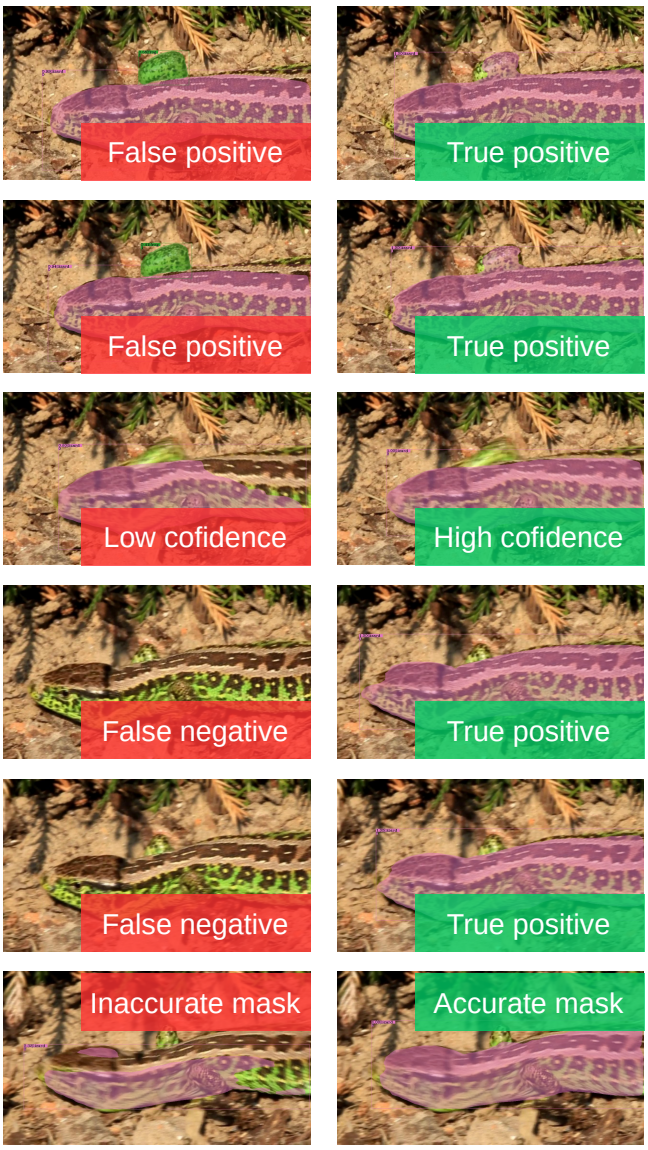

Fig. S3. Typical cases for video instance segmentation. (a) and (b) show the detection results of Swin-T and TDViT-T, respectively. In occlusion and blur scenarios, the performance of Swin-T is drastically decreased, while our TDViT-T still works well.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
2. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
5. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019)