Supplementary Material for SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation

Brendan Duke[1,4*] Abdalla Ahmed[4] Christian Wolf[3] Parham Aarabi[1,4] Graham W. Taylor[2,5]

[1]University of Toronto [2]University of Guelph [3]Université de Lyon, INSA-Lyon, LIRIS

[4]Modiface, Inc. [5]Vector Institute

1. Additional Results

In Figures 1 and 2, we compare SST qualitatively to CFBI [2] and STM [1]. SST produces superior tracking in these challenging sequences, which contain occlusions and disocclusions. The positional encoding in the Transformer representations may enable SST to distinguish similar instances under occlusions, using positional information. Whereas CFBI confuses instances that are far apart, SST remains robust to these nonlocal failures. This further supports the effectiveness of SST's use of positional information.

2. Grid Attention Routing

To demonstrate mathematically information propagation in grid attention we consider, for sake of clarity, a sparse attention function

$$\texttt{SparseAttn}\big(\mathbf{Q},\mathbf{K},\mathbf{V}\big)_{\mathbf{p}} = \mathbf{Q}_{\mathbf{p}}\mathbf{K}_{I_{\mathbf{p}}}^{\intercal}\mathbf{V}_{I_{\mathbf{p}}}. \tag{1}$$

where we replace the softmax by an identity function. Further assume that our query, key, and value all are our video feature tensor, i.e., $\mathbf{Q}=\mathbf{T}$, $\mathbf{K}=\mathbf{T}$, and $\mathbf{V}=\mathbf{T}$. The first layer outputs, for each pixel $\mathbf{p}$,

$$\begin{aligned}\texttt{GridAttn}\big(\mathbf{Q},\mathbf{K},\mathbf{V}\big)_{\mathbf{p}xyt} = &\sum_{i=1}^{W}\big(\mathbf{T}_{xyt}^{\intercal}\mathbf{T}_{iyt}\big)\mathbf{T}_{iyt}+\\ &\sum_{\substack{j=1\\ j\neq y}}^{H}\big(\mathbf{T}_{xyt}^{\intercal}\mathbf{T}_{xjt}\big)\mathbf{T}_{xjt}+\\ &\sum_{\substack{k=1\\ k\neq t}}^{T}\big(\mathbf{T}_{xyt}^{\intercal}\mathbf{T}_{xyk}\big)\mathbf{T}_{xyk}.\end{aligned} \tag{2}$$

We can show that composing three applications of self-attention on $\mathbf{T}$, which we denote for brevity as $\texttt{GridAttn}^3$, produces

$$\begin{aligned}\texttt{GridAttn}^3\big(\mathbf{T},\mathbf{T},\mathbf{T}\big)_{xyt} = \sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{T}&\big(\mathbf{T}_{xyt}^{\intercal}\mathbf{T}_{iyt}\big)\\ &\big(\mathbf{T}_{iyt}^{\intercal}\mathbf{T}_{ijt}\big)\\ &\big(\mathbf{T}_{ijt}^{\intercal}\mathbf{T}_{ijk}\big)\mathbf{T}_{ijk}+\cdots,\end{aligned} \tag{3}$$

where $\cdots$ represents other similar third order terms. We show in Equation 3 that grid attention propagates information along "routes" through the video feature tensor: for a pixel at position (x,y,t) to interact with another pixel at an arbitrary position (i,j,k), interactions must propagate along a "route" through the video feature tensor of pairs of similar pixels. Just as we might give travel directions through a city grid such as "first walk ten blocks North, then walk three blocks East", grid attention interactions must propagate a fixed number of pixels in the X, Y and T directions, in some order, before connecting the interaction source pixel with its target pixel.

Consider what happens if we replace the value $\mathbf{T}_{ijk}$ returned by the inner cross-attention in Equation 3 with a foreground mask value m_{ijk}. We see that the output routes reference mask values m_{ijk} over paths of feature vectors in the video tensor $\mathbf{T}$ that transitively correspond to reference features $\mathbf{T}_{ijk}$.

*Corresponding Author: brendanw.duke@gmail.com

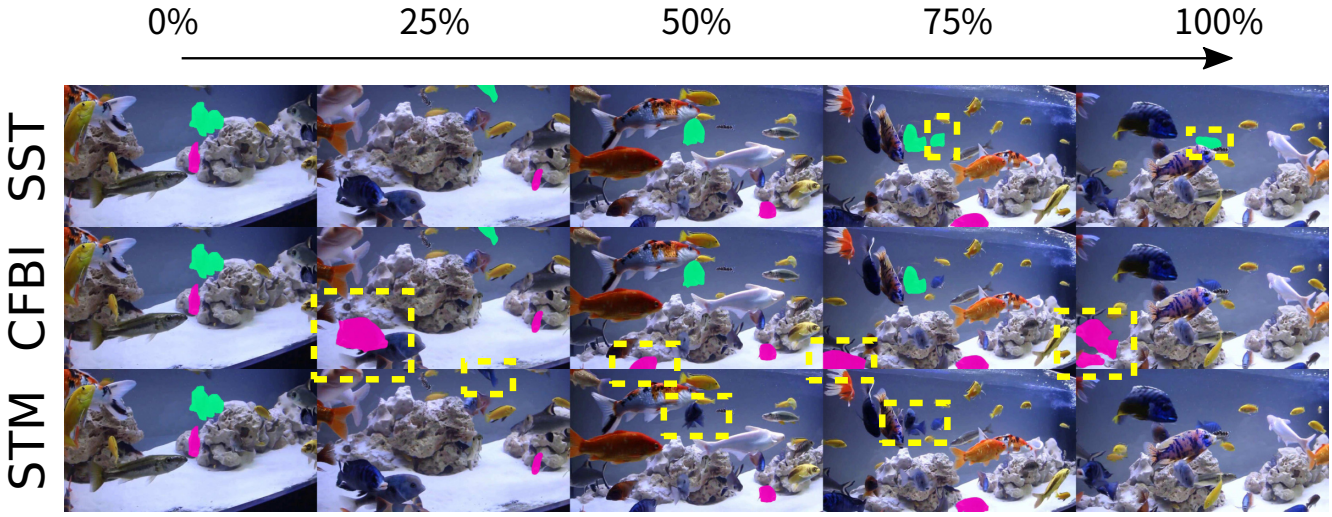

Figure 1: Fish tank. This challenging YouTube-VOS 2019 validation set example contains many occlusions and disocclusions by similar-looking instances of the same fish class. SST makes relatively few errors relatively later in the sequence when compared with CFBI [2] or STM [1]. For clarity we labeled errors with yellow dotted boxes (best viewed digitally, with zoom and in colour).

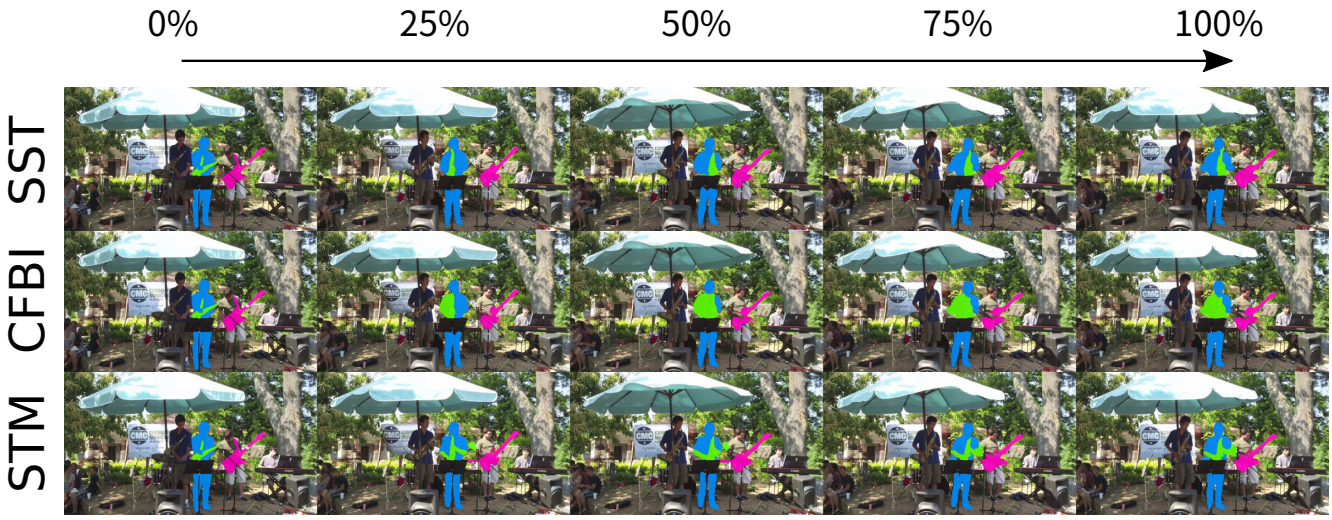

Figure 2: Jazz band. In this YouTube-VOS 2019 validation set example the saxophone player self-occludes and disoccludes their saxophone while playing. SST maintains the correct saxophone segmentation throughout the sequence. In contrast, CFBI [2] and STM [1] confuse the saxophone with the saxophone player's upper body after disocclusion.

References

[1] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. *ICCV*, 2019.

[2] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *Proceedings of the European Conference on Computer Vision*, 2020.