Supplementary: BATMAN: Bilateral Attention Transformer in Motion-Appearance Neighboring Space for Video Object Segmentation

Ye Yu[1], Jialin Yuan[2], Gaurav Mittal[1], Li Fuxin[2], and Mei Chen[1]

[1] Microsoft
{yu.ye,gaurav.mittal,mei.chen}@microsoft.com
[2] Oregon State University
{yuanjial,lif}@oregonstate.edu

1 Local window size in the bilateral space

We analyze the local window size in the bilateral space (W_b) of the bilateral attention module and its impact on VOS performance. Table 1 shows the importance of a proper W_b for bilateral attention and VOS. When the W_b is too small, relevant tokens are missed out in the bilateral attention, which decreases the VOS performance. On the other hand, if we have too large W_b size, more irrelevant tokens will be included in the bilateral attention computation (similar to the spatial local attention), which also decreases the VOS performance.

Table 1: Ablation on bilateral local window (W_b) size. The experiment is on DAVIS 2017 validation split. W_b size needs to be carefully chosen for optimal VOS performance

W_b size	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
40	84.2	81.3	87.0
60	84.7	81.8	87.5
84	**86.2**	**83.2**	**89.3**
112	84.9	81.9	87.9

2 Additional segmentation visualizations

Below we provide additional segmentation visualizations on DAVIS 2017 (Fig. 1) and Youtube-VOS 2019 (Fig. 2). On both datasets, the bilateral attention segments object better compared to the spatial local attention. The bilateral attention can distinguish the objects with a similar visual appearance especially when they have different motions, which the conventional spatial local attention struggles to segment.

Fig. 1: Additional qualitative results on DAVIS 2017. Compared to spatial local attention, bilateral attention segments object better especially when background shares similar appearance with the target object

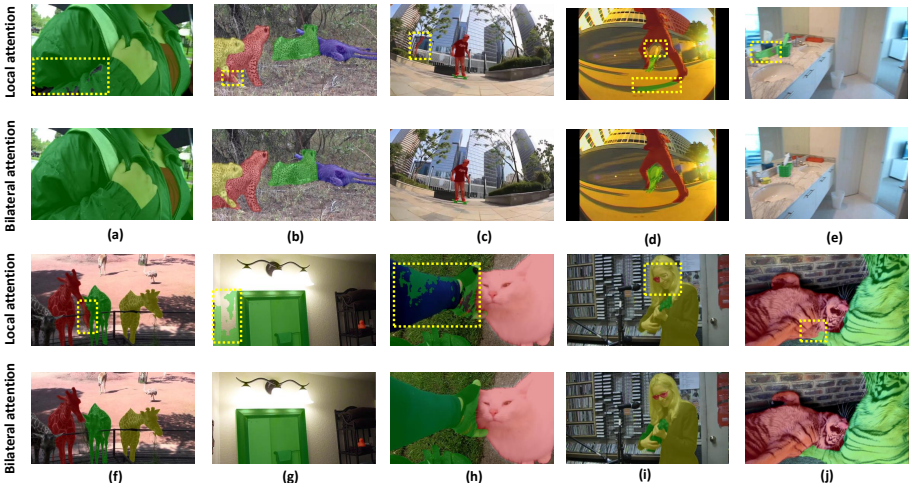

Fig. 2: Additional qualitative results on Youtube-VOS 2019. Compared to spatial local attention, bilateral attention segments object better especially when the object has a salient motion