We address the problem of semi-supervised video object segmentation (VOS), where the masks of objects of interests are given in the first frame of an input vid