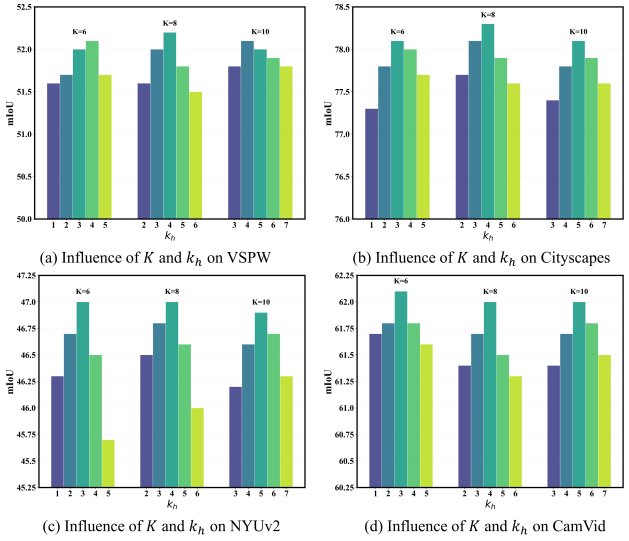
RS-SSM: Refining Forgotten Specifics in State Space Model for Video Semantic Segmentation

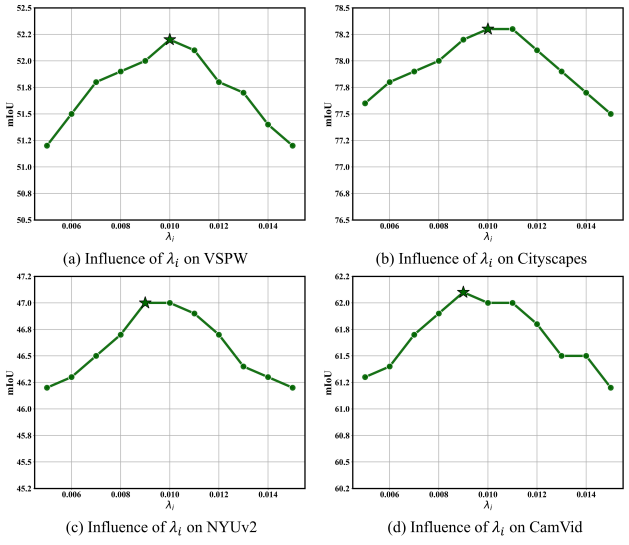
Supplementary Material



(a) Influence of K and k_h on VSPW (b) Influence of K and k_h on Cityscapes

(c) Influence of K and k_h on NYUv2 (d) Influence of K and k_h on CamVid

Figure 5. Influence of K and k_h mentioned in Eq 13.



(a) Influence of λ_i on VSPW (b) Influence of λ_i on Cityscapes

(c) Influence of λ_i on NYUv2 (d) Influence of λ_i on CamVid

Figure 6. Influence of λ_i mentioned in Eq 20.

7. Influence of Hyper-parameters

We conducted ablation experiments on the VPSW, Cityscapes, NYUv2 and CamVid dataset to evaluate the impact of several hyperparameters, including the total number of frequency bands K, the number of high-frequency bands k_h as mentioned in Eq 13, and the weight λ_i of the channel information loss as mentioned in Eq 20. The results are shown in Fig 5 and Fig 6. On the Cityscapes dataset, we use MiT-B1 as the backbone, while on other datasets, we employ Swin-S as the backbone.

We observe that K has minimal impact on mIoU, with $K = 8$ performing slightly better. For k_h, mIoU decreases when it exceeds 50% of K, as aggregating excessive frequency bands introduces redundant information that impairs spatiotemporal specifics modeling. Conversely, a low proportion also degrades performance due to insufficient spatiotemporal specifics distribution statistics. In our implementation, the K is set to 8 and the k_h is set to 4.

For λ_i, our method achieves the best performance when we set it to 0.01. When λ_i decreases, performance drops as the loss term fails to align channel distributions across samples. However, large values also reduce mIoU by interfering with the primary segmentation task optimization. Therefore, we set the default value of λ_i to 0.01.

8. Visualization of More Cases

8.1. More Visualizations of the Updating Gate

As shown in Eq 2, the updating gate $\overline{\mathbf{B}}_d$ in SSM model is influenced by the forgetting gate, reflecting the amount of new information introduced at each time step. To verify the effectiveness on specifics refining of our proposed RS-SSM, we visualize the value of updating gate. As shown in Figure 7, our RS-SSM can effectively refine the spatiotemporal specifics forgotten during the state space compression. This is attributed to our designed CwAP and FGIR modules, which can adaptively invert and refine the forgetting gate in SSM, demonstrating the effectiveness of our design.

8.2. More Visualizations of Segmentation Results

As shown in Fig 8, we visualize several segmentation results on the VSPW dataset. Compared to the existing SSM-based method TV3S, our RS-SSM produces more accurate and detailed segmentation results, especially for small objects and object boundaries, as well as scenarios involving semantic object displacement and mutual occlusion. This is because our proposed CwAP and FGIR modules can effectively refine the spatiotemporal specifics forgotten in SSM, thereby maintaining the model's modeling capability for specific information across video streams, which further validates the effectiveness of our method.

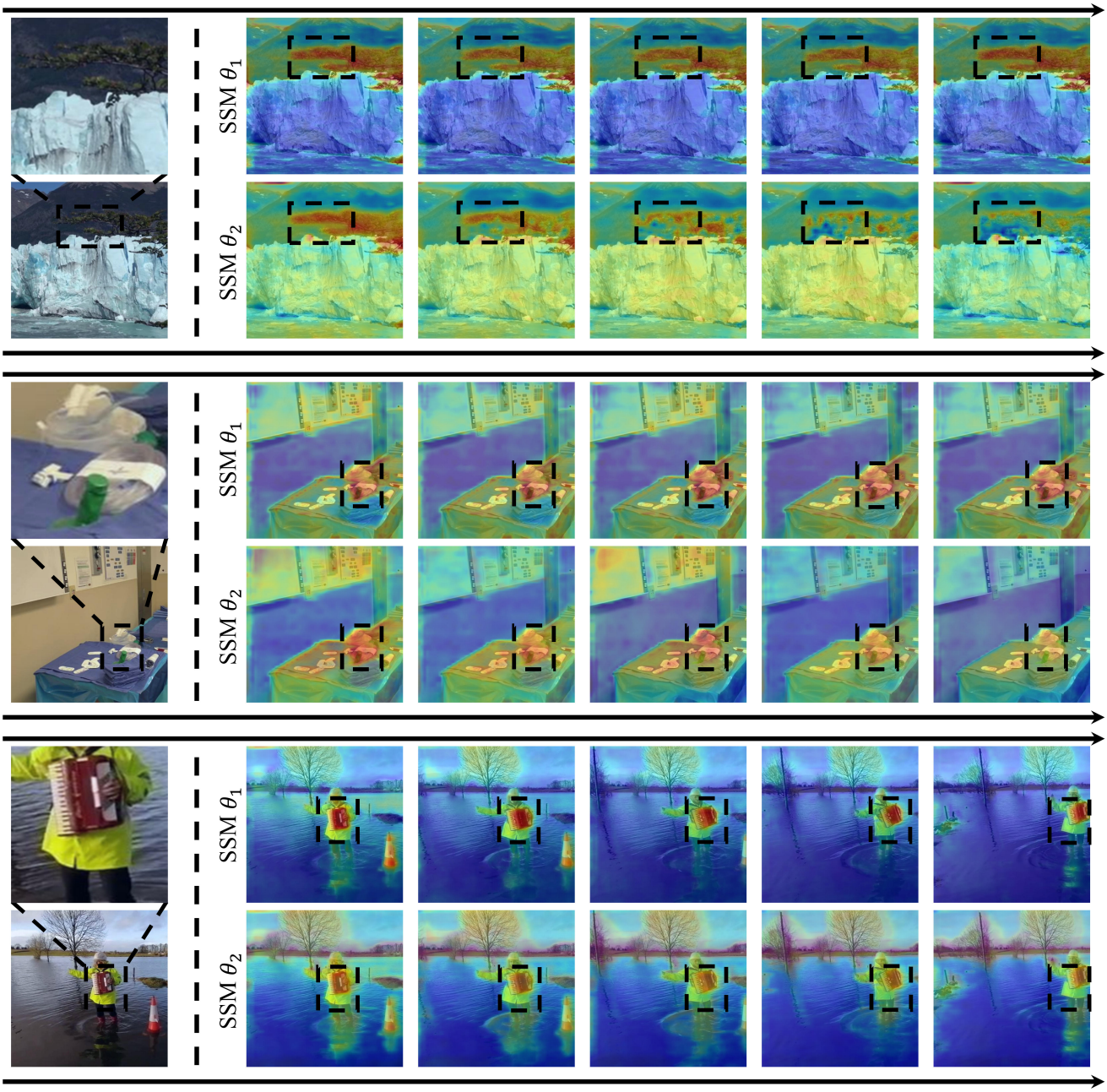


Figure 7. Visualization of the updating gate. The bottom row visualization reveals that the vanilla SSM θ_2 suffers from information loss of specifics during the state space compression. Conversely, as illustrated in the top row, the SSM θ_1 effectively refines these forgotten specifics after inverting and refining the forgetting gate.

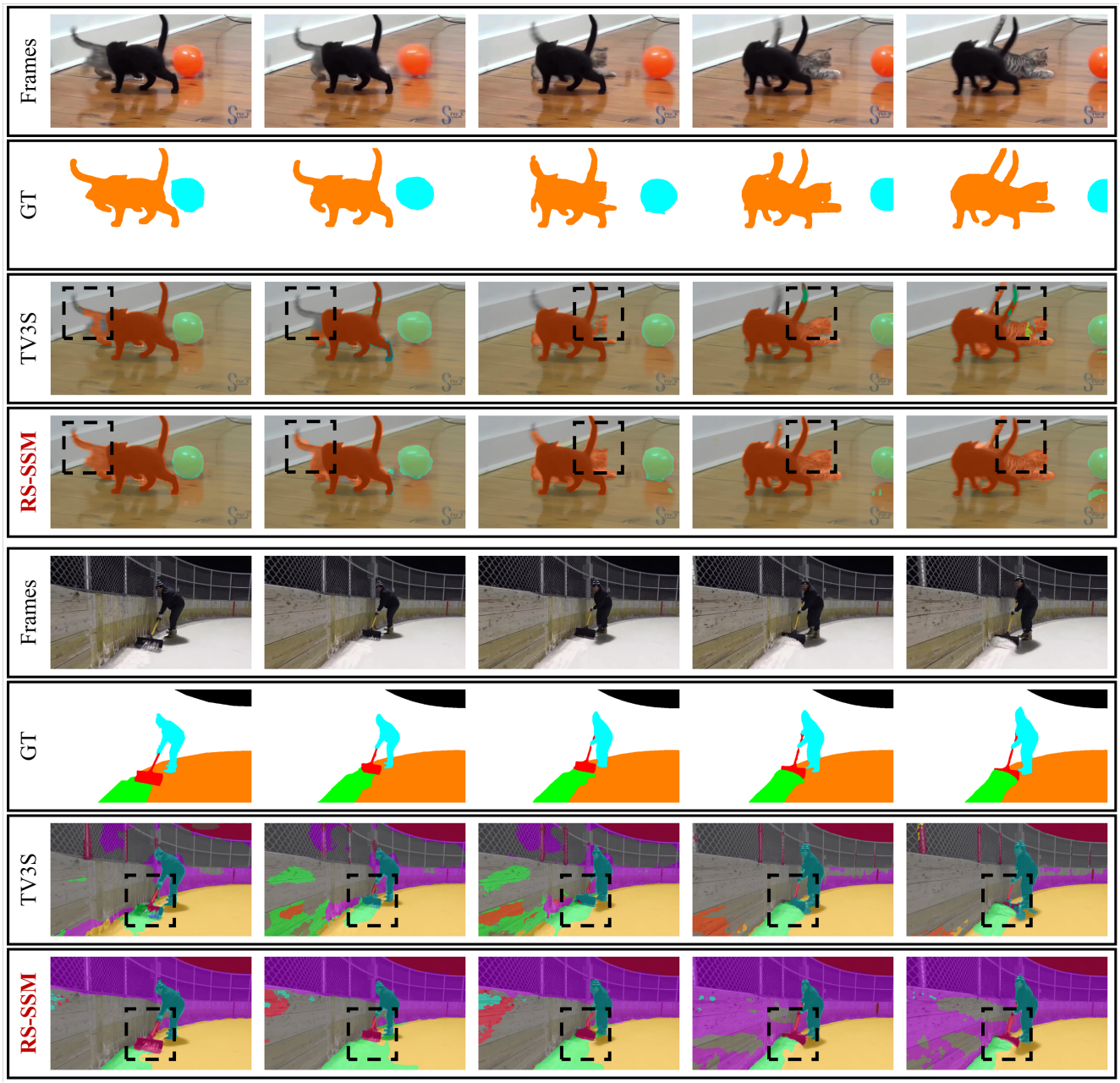


Figure 8. Visualization of segmentation results on VSPW dataset. Compared to the existing SSM-based method TV3S, our RS-SSM produces more accurate and detailed segmentation results by effectively refining specific information in videos.