GeoMotion: Rethinking Motion Segmentation via Latent 4D Geometry

Supplementary Material

Algorithm 1: Feature Aggregation Module

Input: $F_{\text{geo}}^{\text{low}}$: low-level geometry features $[B{*}N,\ hw,\ 2C]$
$F_{\text{geo}}^{\text{high}}$: high-level geometry features $[B{*}N,\ hw,\ 2C]$
F_{cam}: camera features $[B{*}N,\ hw,\ 512]$
flow: optical flow map $[B,\ N,\ H,\ W]$
C: 1024
Output: F_{fuse}: fused features $[B{*}N,\ hw,\ 2048]$

```
// 1) Concatenate multi-level geometry features
```
$F_{\text{geo}} \leftarrow \texttt{Concat}(F_{\text{geo}}^{\text{low}}, F_{\text{geo}}^{\text{high}})$;
// $[B{*}N,\ hw,\ 4C]$
```
// 2) Project geometry features to 2048-d
```
$F_{\text{geo}} \leftarrow \texttt{Linear+ReLU}(F_{\text{geo}})$;
// $[B{*}N,\ hw,\ 2048]$
```
// 3) Encode optical flow and convert to patch tokens
```
$F_{\text{flow}} \leftarrow \texttt{BilinearDown}(\texttt{CNN}(\text{flow}), h,\ w)$;
// $[B{*}N,\ hw,\ 128]$
```
// 4) Fuse geometry, flow, and camera features
```
$F_{\text{cat}} \leftarrow \texttt{Concat}(F_{\text{geo}}, F_{\text{flow}}, F_{\text{cam}})$;
// $[B{*}N,\ hw,\ 2688]$
```
// 5) Final projection
```
$F_{\text{fuse}} \leftarrow \texttt{Linear}(F_{\text{cat}})$; // $[B{*}N,\ hw,\ 2048]$

return F_{fuse}

A. Architecture Details

Algorithm 1 presents the detailed pseudo-code of the proposed Feature Aggregation Module, providing implementation-level clarification of the fusion and projection steps.

B. Ablation for SAM2

The results of SAM2 ablation are shown in Tab. 5 and Fig. 6. SAM2 mainly enhances boundary quality, and our core method remains effective without it. The raw output of the model still surpasses other refined methods like Easi3R w/SAM2.

Method	JM↑		JR↑		FM↑	
	w/o SAM	w/ SAM	w/o SAM	w/ SAM	w/o SAM	w/ SAM
Easi3R$_{\text{dust3r}}$ [3]	46.86	60.1	50.54	65.3	39.06	–
Easi3R$_{\text{monst3r}}$ [3]	54.75	67.9	66.16	76.1	44.09	–
MonST3R [55]	38.07	56.4	36.05	59.6	48.24	–
DAS3R [49]	44.51	57.4	43.95	61.3	46.71	–
VGGT4D [11]	56.45	–	65.62	–	51.09	–
OCLR-flow [46]	69.90	–	–	–	–	–
ABR [48]	74.60	–	–	–	–	–
Ours	**75.38**	**81.13**	**87.19**	**92.63**	72.29	81.82

Table 5. Ablation for SAM2 on DAVIS-2017.

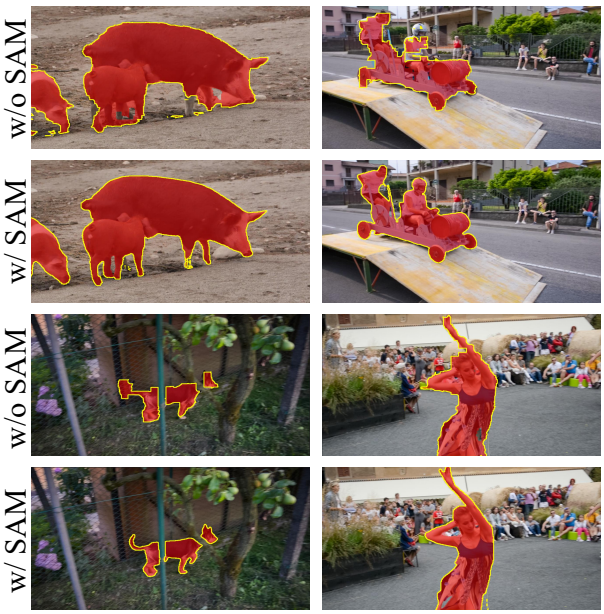

Figure 6. Qualitative ablation results for SAM2.

C. Qualitative Comparison with Reconstruction Methods

Figure 7 shows qualitative comparisons with reconstruction-based methods, including Easi3R [3] and VGGT4D [11], using raw predictions without SAM2 refinement. GeoMotion produces cleaner and more compact motion masks with fewer background false positives. In contrast, reconstruction-based approaches may over-segment background regions or generate fragmented masks under weak motion cues. Our geometry-, flow-, and camera-aware fusion results in more stable and coherent object-level predictions.

Figure 7. Compare with Easi3R and VGGT4D without SAM2.

D. More Visualizations on Dynamic Scenes

We provide additional qualitative visualizations in Fig. 8 to further illustrate the robustness of GeoMotion. Seven representative sequences are selected from the DAVIS benchmark: the first three depict scenes with a single moving object, while the remaining four contain multiple moving objects. Across diverse challenging scenarios such as motion blur, heavy occlusion, fast object motion, large camera movement, and strong appearance similarity, GeoMotion consistently produces accurate and temporally stable dynamic masks. These results indicate that the geometry-driven representation effectively captures motion cues even when conventional signals are unreliable.

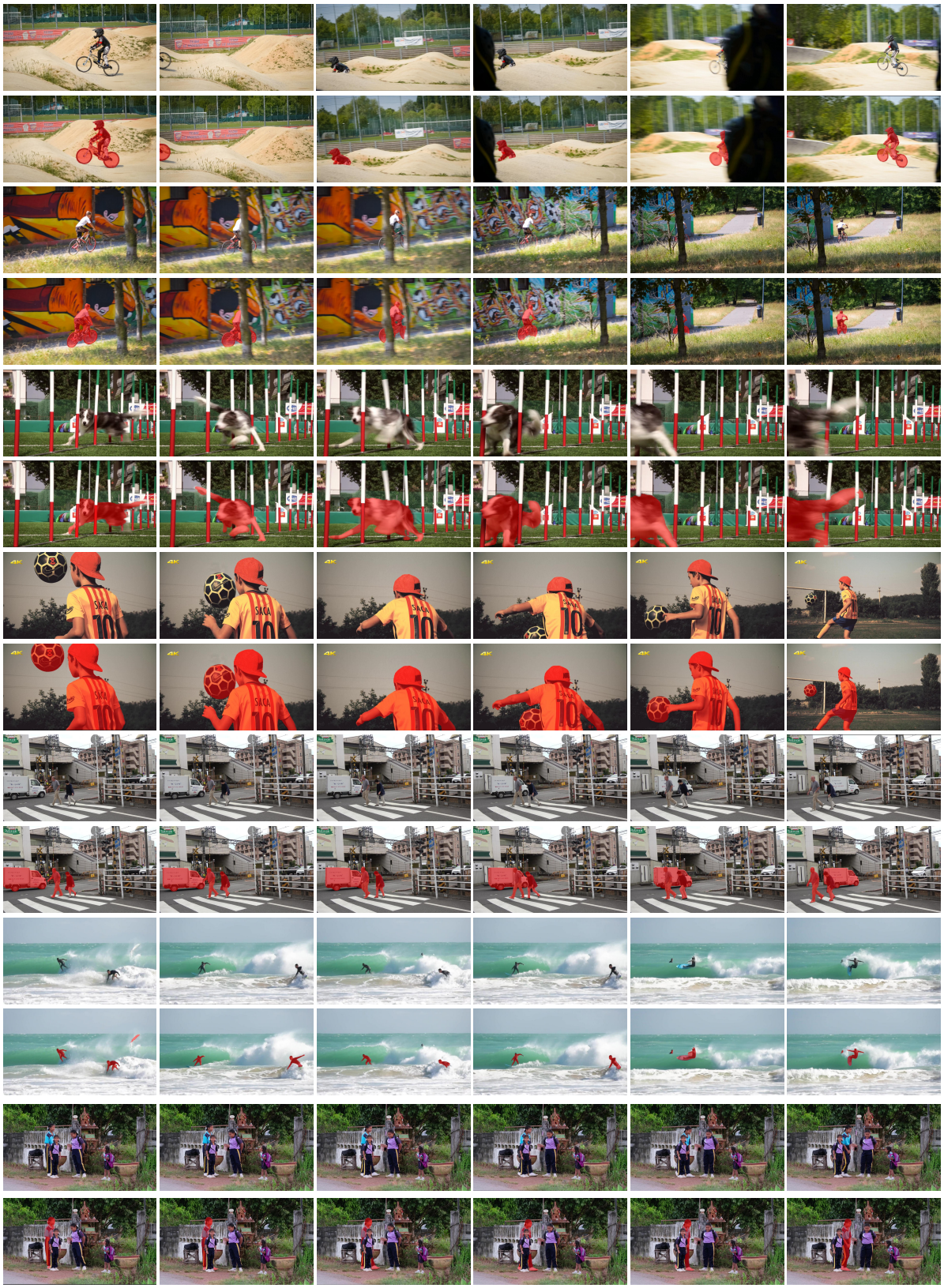

Figure 8. **More visual examples of dynamic masks predicted by GeoMotion on the DAVIS benchmark.** Odd rows show the RGB input frames, while even rows present the corresponding predicted dynamic masks.