Supplementary Material
SeqFormer: Sequential Transformer for Video Instance Segmentation

Junfeng Wu[1], Yi Jiang[2], Song Bai[2], Wenqing Zhang[1], and Xiang Bai[1†]

[1] Huazhong University of Science and Technology [2] Bytedance

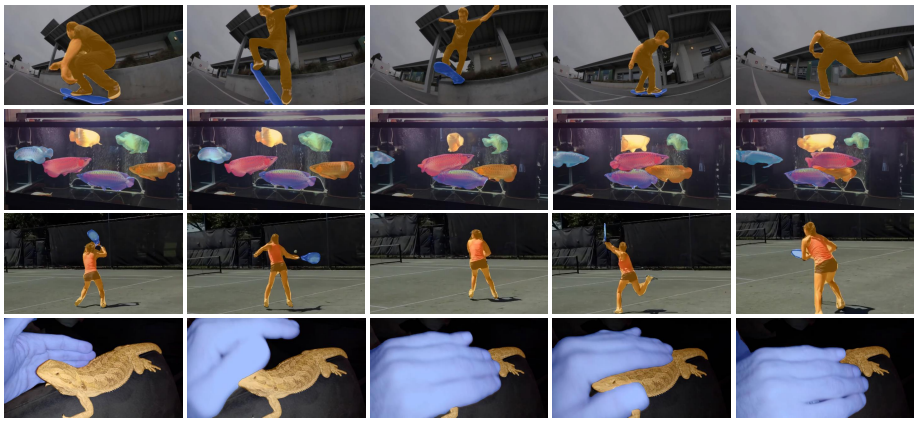

Fig. 1. Visualization of SeqFormer on the YouTube-VIS 2019 validation dataset. The first row shows the instances with various poses. The second row shows the case of a lot of similar instances that are close together with overlapping. The third row shows the situation where an instance reappears after being occluded while in motion. The last row shows an instance severely occluded by the other instance. The same colors depict the mask sequences of the same instances

1 Visualization

In Fig. 1, we visualize the results of SeqFormer with four challenging cases. It can be seen that SeqFormer can handle these situations well. In Fig. 2, we show more qualitative results of the intermediate attention of transformer decoder. Since the same initial instance query is used to predict sampling points for each frame in the first decoder layer (Eq.1), the distribution of sampling points on each frame is the same in Fig. 2 (a) and (d). After that, the initial instance query is decomposed into frame-level box queries that are kept and maintained independently on each frame. Starting from the second layer of the SeqFormer decoder, the box query is used to predict the sampling points of the

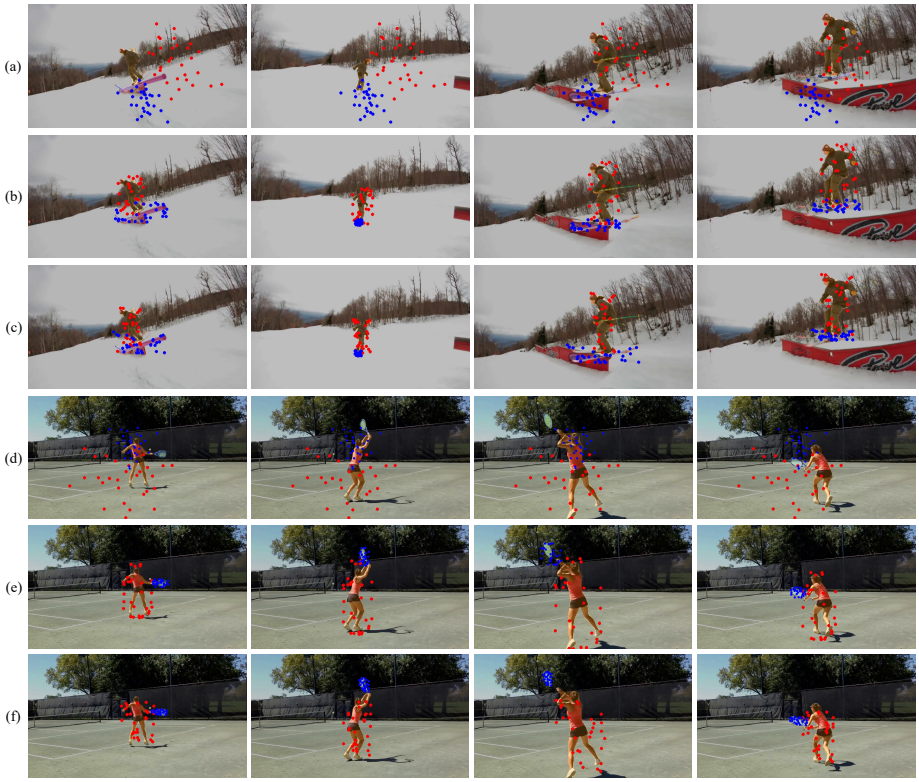

Fig. 2. Visualization of attention. We draw the sampling points that the deformable attention attends to. The four frames in each row are from the same video. Each sampling point is marked as a filled circle whose color indicates its corresponding instance identity. (a) and (d) show the sampling points from the first decoder layer. (b) and (e) show the sampling points from the second decoder layer. (c) and (f) show the sampling points from the last decoder layer.

current frame, and the sampled features are used to refine the box query for the next decoder layer. By doing so, SeqFormer attends to different spatial locations following the motion of the instance in a coarse-to-fine manner.

2 Aggregation of Temporal Information

SeqFormer is able to attend to different spatial locations following the motion of the instance. The aligned features are aggregated into an instance query to generate a video-level instance representation. However, an instance may not appear in every frame due to occlusion and camera motion. The features from frames without instance are useless or even harmful. To address this, SeqFormer aggregates temporal features in a weighted manner, where the weights

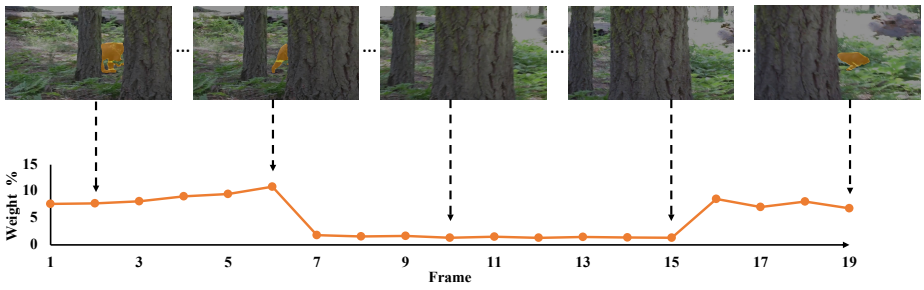

Fig. 3. Visualization of the normalized softmax weights and the corresponding frames.

are learned upon the box queries in Eq.3. We visualize the learned weights and the corresponding frames in Fig. 3. It can be seen that the features from frames without instance have lower weights.

3 Qualitative Comparisons

We provide some qualitative comparisons with other methods in Fig. 4, the mask predictions of SeqFormer are more stable over time. More video results and comparisons can be found in the rest of the supplementary material.

4 Clip Matching

Our model can be extended to per-clip model through clip matching algorithm to handle long videos. Specifically, we divide long videos into clips with overlapping frames, and match clip-level instance masks by calculate the matching scores which are space-time soft IoU of overlapping frames, following IFC. We evaluate our method by varying the length of clips in Table 1. Our method still achieves competitive performance but slightly worse when evaluating in the clip-wise manner. This manner makes our method handle very long videos with limited computational resources and have a wider range of application scenarios.

Table 1. Evaluating SeqFormer in a clip-wise manner.

Clip Length	Whole	20	15	10
AP	45.1	44.8	43.6	42.0

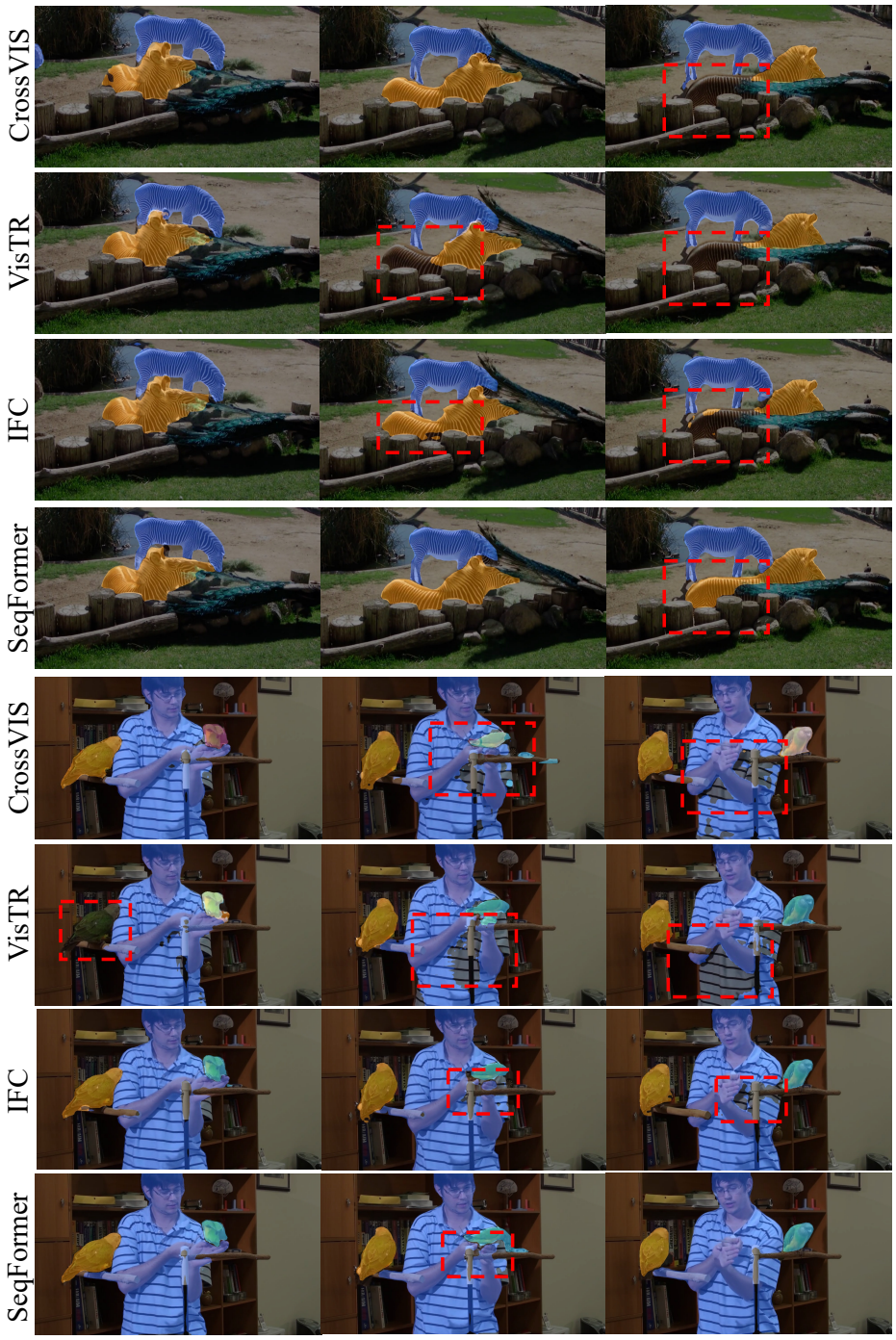

Fig. 4. Qualitative comparisons with other methods on YouTube-VIS 2019. All methods use ResNet-50 backbone. The three frames in each row are from the same video. The mask predictions of SeqFormer are more stable over time. Best viewed in color.