Supplementary Materials for SRNet: Spatial Relation Network for Efficient Single-stage Instance Segmentation in Videos

Xiaowen Ying
Lehigh University
xiy517@lehigh.edu

Xin Li
Lehigh University
xincoder@gmail.com

Mooi Choo Chuah
Lehigh University
chuah@cse.lehigh.edu

A ADDITIONAL ABLATION STUDIES

Center Selection for Tracking. During post-processing, we use the predicted tracking vectors to associate the instance center with its location at the previous frame. In this process, it is possible to select different types of instance centers for associations. Table 1 summarizes the results of different choices. The Max Objectiveness is the location that has the maximum objectiveness score, corresponding to c'_n in our formulation. The Center of Cluster corresponds to c_n obtained from equation (4). The Center of BBox denotes the bounding box center. It is worth mentioning that our approach is box-free so the bounding box center is calculated by first converting the instance mask to its corresponding bounding box. The Center of Mass is calculated w.r.t the instance mask. Among all these choices, we found that the Center of Mass is more stable compared to other choices. For example, the bounding box center could sometimes be outside of the object depending on the object's shape. The Center of Cluster is very close to the Center of Mass in most cases; however, since it is learned by the model, the model occasionally adjusts its location to reduce the difficulty of clustering pixels. An example scenario is when two objects are closed to each other in the scene, the model may put their cluster centers to the side and away from each other so it is easier to separate them.

Impact of Matching Algorithms. We compare the Greedy Assignment used in our post-processing with another popular choice – the Hungarian Assignment. Our experimental results in Table 2 show that the Greedy Assignment is not only simple and more efficient but also performs better in our case. The difference is that the Greedy Assignment simply selects the best available candidate during each iteration, while the Hungarian algorithm tries to minimize the total assignment cost. One reason that causes this performance gap is that our model is learned in a class-agnostic fashion and it sometimes produces instance candidates for irrelevant background objects if their objectiveness score is high enough. The predicted tracking vectors at those background locations are not very stable which could point those candidates to a random location. In such cases, minimizing the total assignment cost could easily cause ID switches. We leave further investigations to future works.

Selected Center	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
Max Objectiveness	40.0	38.5	41.4
Center of Cluster	55.3	53.7	56.9
Center of BBox	52.4	50.4	54.4
Center of Mass	59.7	58.2	61.3

Table 1: Impact of using different center for tracking.

Algorithm	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
Hungarian Assignment	48.4	46.4	50.5
Greedy Assignment	59.7	58.2	61.3

Table 2: Impact of using different assignment algorithm.

B ADDITIONAL SPEED ANALYSIS

Inference Time of Each Component. We provide further analysis of the inference time of each component of our framework in Table 3. The results further demonstrate the efficiency of our core designs especially the decoder and post-processing parts.

Mode	BS	FE	DE	CL	TR	FPS
Offline	16	19.6	5.7	3.2	0.3	35
Offline	8	22.3	5.7	3.2	0.3	32
Streaming	1	30.1	5.7	3.2	0.3	25

Table 3: Inference Time (ms) of Each Components. BS: Batch Size for parallel feature extraction. FE: Feature Extraction. DE: Decoder. CL: Clustering. TR: Tracking Association.

Offline Processing V.S. Streaming. Table 3 also shows that the time spends on feature extraction is a bottleneck of our inference speed. We purposely design our framework such that the feature extraction part is temporal-independent so it can be accelerated using parallel processing in offline mode. This is similar to the practice of Stem-Seg [1] and VisTR [43] which equivalently process multiple frames in a single pass (16 frames for StemSeg and 36 frames for VisTR). It is worth mentioning that these methods can only perform offline processing while our framework is flexible to operate in both offline and streaming modes. Even in strictly streaming mode, our inference speed is still significantly faster than existing methods compared in our main paper.

C ADDITIONAL QUALITATIVE ANALYSIS

In this section, we provide more qualitative results on both DAVIS-2019 and Youtube-VIS, illustrated in Figure 1 and 2, respectively. These examples include challenging scenes such as objects with rapidly changing shapes, overlapping objects, crowded streets, and objects that are partially occluded. We can see that our method can handle a variety of different scenarios well.

D FAILURE CASES

Our method only considers short-term temporal information which allows us to perform high-speed online processing. However, such a design naturally comes with the limitation on the ability to associate long-range tracks. For example, if an object is occluded by other objects in a certain frame and reappears later, it may end up with a new track ID. We provide example failure cases as illustrated in Figure 3. This issue could potentially be alleviated by incorporating an extra Re-ID module at the cost of longer processing time, which we leave to our future works.

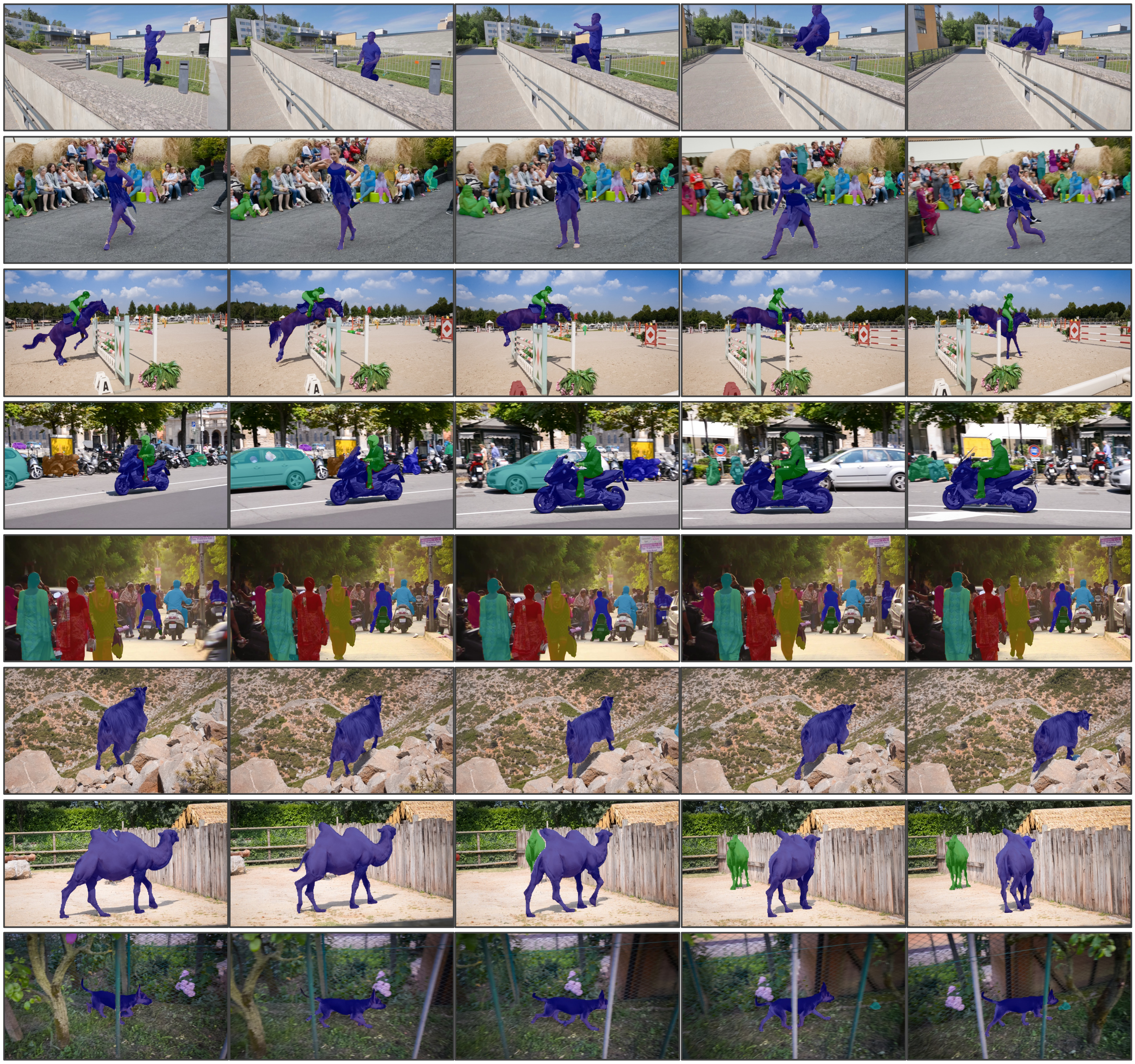

Figure 1: More qualitative results on DAVIS-2019 Validation Set. Our framework is able to produce satisfying results under diverse scenarios. The first two rows demonstrate the ability of our network to handle objects that experience rapid changes in shape (parkour and dancing). The third and fourth rows show two challenging cases where two objects are overlapped with each other (horse-riding and bike-riding). Our results clearly separate the two objects from each other, especially for the challenging legs area. The fifth row shows a scene with crowded people and our approach is able to produce accurate masks and keep tracks of many objects simultaneously. The last row shows a challenging scene where the dog is frequently occluded by pipes in the foreground, and our framework is still able to segment the entire dog and keep track of its ID.

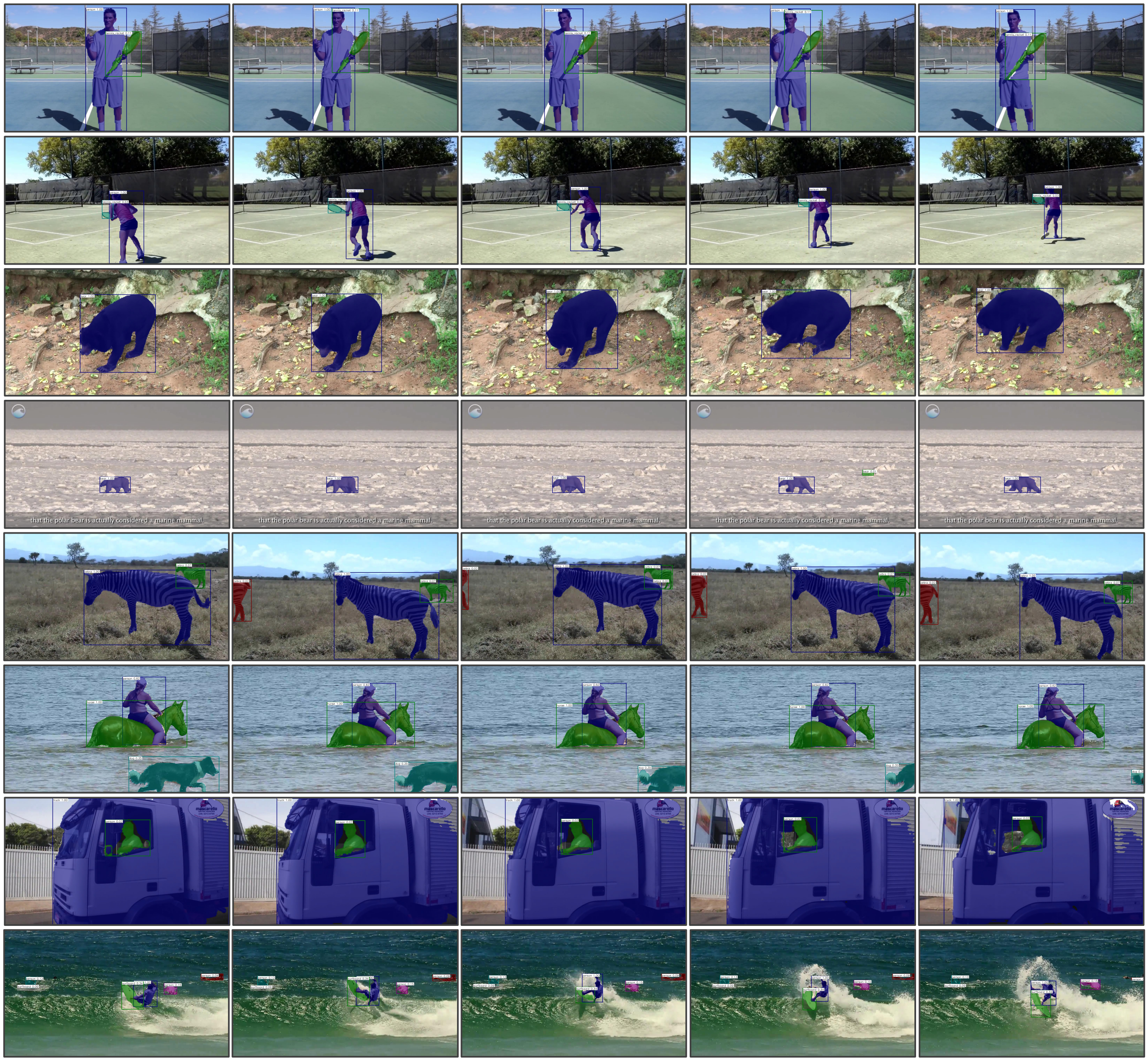
that the polar bear is actually considered a marine mammal
that the polar bear is actually considered a marine mammal
that the polar bear is actually considered a marine mammal
that the polar bear is actually considered a marine mammal
that the polar bear is actually considered a marine mammal

Figure 2: More qualitative results on Youtube-VIS Validation Set. We show that by simply adding a lightweight category head our approach can also handle the VIS task well. The first two rows demonstrate the scenarios of humans holding objects in both still and moving cases. Our approach is able to capture both salient objects (humans) and fine details (tennis brackets). The third to fifth rows are scenarios with a single animal, an animal with colors similar to the background (polar bear), and multiple animals, respectively. The sixth and seventh rows are scenarios where objects have large overlaps, especially the person sitting in the trunk in the seventh row. The last row shows a challenging scenario where everything in the scene is rapidly changing including the object movements and backgrounds. Our framework consistently segments and tracks multiple surfers and their surfboards.

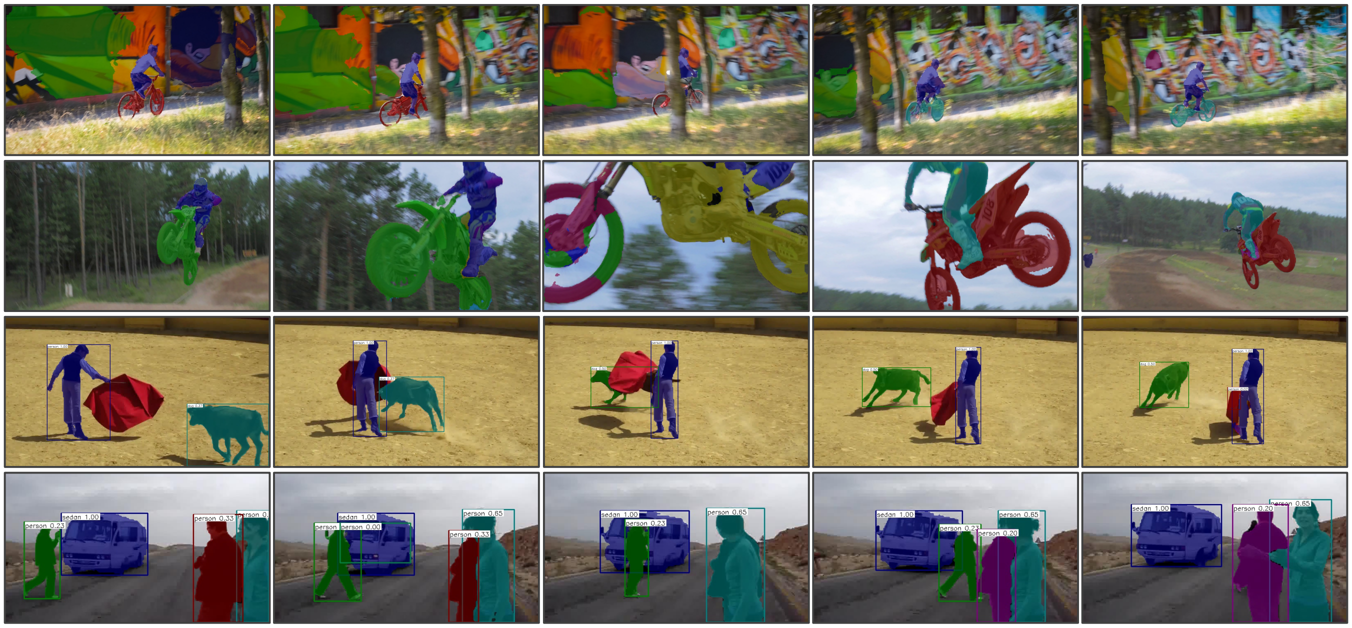

Figure 3: Example failure cases. The first two rows are sampled from the DAVIS-2019 Validation Set and the last two rows are sampled from the YoutubeVIS validation set. These failure cases mostly correspond to the limitation of our approach to connect long-range tracks. The first row shows an example where we miss the bike in the third frame because it is mostly occluded by the tree. When we detected it again we assign it with a new track ID (note that the biker is successfully tracked). The second row shows a similar example where the rider is out of the camera in the third frame and ID-switches happen to both objects in subsequent frames. The third row shows a bullfight scene where the ID switch happens to the bull when it runs through the cape. Also, the model misclassified the bull as a dog. In the last row, the two people on the right are too close to each other and the network merged them into one instance in the third frame thus a new track ID has been assigned when they reappeared.