Supplementary: Sync from the Sea: Retrieving Alignable Videos from Large-Scale Datasets

Ishan Rajendrakumar Dave[1*], Fabian Caba Heilbron[2], Mubarak Shah[1], and Simon Jenni[2]

[1] Center for Research in Computer Vision, University of Central Florida, USA
[2] Adobe Research, USA
ishandave@ucf.edu, caba@adobe.com, shah@crcv.ucf.edu, jenni@adobe.com
https://daveishan.github.io/avr-webpage/

A Detailed Evaluation for Alignable Video Retrieval

We provide an expanded table of Table-1 of our main paper in Table 1 here. For all the combinations of alignment features and candidate retrievals, we additionally include the cases of using DTW re-ranking and non-contextualized features. We make the following observations:

- Within comparable 'Alignment Features' and 'Reranking Metrics', we observe significant improvements when employing our novel feature contextualization approach. For example, compare `Row (a1)` with `Row (d1)`, `Row (b1)` with `Row (e1)`, and `Row (c1)` with `Row (f1)`, etc.
- Within comparable 'Alignment Features' and 'Context', our proposed DRAQ-based reranking scheme outperforms both the DTW-based reranking and the absence of reranking. For instance, `Row (c1)` with DRAQ surpasses `Row (a1)` with no reranking, and `Row (b1)` with DTW reranking.

B Qualitative Results

Qualitative AVR examples of retrieved and aligned video pairs on Kinetics700 are provided in the Supplementary videos. Our qualitative results underscore that our method can effectively achieve alignment within a large-scale dataset, where the top video displays the query and the bottom video showcases the best alignable match retrieved, where both videos are warped with the optimal alignment path P_{DTW}.

For instance, our method precisely matches action phases across various scenarios. This includes watermelon cutting techniques in `2.mp4`, catch-throw-stance sequences in `9.mp4`, and the procedural steps within a manufacturing process video in `16.mp4`. Noteworthy is our method's applicability to a range of general action videos, such as interactions with dolphins in `6.mp4`, the rotating blade of a coffee machine in `5.mp4`, and fishing activities in `15.mp4`.

* Majority of work done as an intern at Adobe Research, USA

	Alignment Features	Context	Reranking Metric	PennAction ↻ FPE	PennAction ↻ CPE	Penn ⇄ UCF FPE	Penn ⇄ UCF CPE	Kinetics ↻ FPE	Kinetics ↻ CPE
Video Candidates obtained through NMS [2] Retrieval									
(a1)	BYOL [3]	✗	-	125.4	3.98	124.5	519.87	564.5	5.06
(b1)			DTW	333.7	7.95	259.8	978.77	1198.4	19.2
(c1)			DRAQ	19.3	0.75	72.1	283.51	2.3	0.24
(d1)		✓	-	0.54	0.4	121.1	105.01	13	1.03
(e1)			DTW	0.78	0.78	206.6	407.08	16.8	2.11
(f1)			DRAQ	0.24	0.13	50.6	11.03	0.32	0.09
(g1)	CARL [1]	✗	-	99.8	2.34	19.2	24.35	40.3	0.68
(h1)			DTW	98.6	1.91	11.3	25.09	12.2	0.34
(i1)			DRAQ	21.1	0.71	21.7	10.62	16	1.12
(j1)		✓	-	90.3	2.38	18.7	28.49	23.5	0.45
(k1)			DTW	78.3	1.68	9.5	15.27	9.6	0.46
(l1)			DRAQ	24.3	0.74	5.2	5.87	2.3	0.08
(m1)	NMS [2]	✗	-	72.9	2.32	37.2	31.9	40	1.06
(n1)			DTW	160.9	3.72	51.8	26	331.1	6.22
(o1)			DRAQ	83.37	0.79	23.5	7.82	0.64	0.21
(p1)		✓	-	13.4	1.32	5.5	22.22	22.7	0.86
(q1)			DTW	14.5	1.86	6.7	29.77	65	1.95
(r1)			DRAQ	9.5	0.2	4.8	5.89	0.46	0
Video Candidates obtained through Oracle Retrieval									
(a2)	BYOL [3]	✗	-	197.5	5.2	-	-	909.2	8.51
(b2)			DTW	79.5	1.19	-	-	980.8	15.14
(c2)			DRAQ	74.5	0.85	-	-	6	0.04
(d2)		✓	-	50.4	4.14	-	-	7.6	0.62
(e2)			DTW	12.7	0.84	-	-	13.4	2.1
(f2)			DRAQ	7.5	0.53	-	-	0.3	0.05
(g2)	CARL [1]	✗	-	41.3	1.46	-	-	47.4	0.41
(h2)			DTW	26.6	1.11	-	-	4.3	0.16
(i2)			DRAQ	31.6	0.41	-	-	2.1	0.19
(j2)		✓	-	23.4	1.34	-	-	36.4	1.04
(k2)			DTW	12.5	0.93	-	-	4.4	0.12
(l2)			DRAQ	11.2	0.36	-	-	1.7	0.14
(m2)	NMS [2]	✗	-	88.1	3.58	-	-	134.6	5.16
(n2)			DTW	64.2	2.97	-	-	157.2	3.18
(o2)			DRAQ	29.5	0.87	-	-	0.2	0
(p2)		✓	-	24.7	1.7	-	-	35.3	1.08
(q2)			DTW	13.3	1.76	-	-	23	0.18
(r2)			DRAQ	7.8	0.33	-	-	0.3	0.01

Table 1: AVR Evaluation. We report additional results without the proposed contextualized features and with different re-ranking schemes.

C Discussion

C.1 Computation Cost, Scalability and Storage

The first stage of our approach involves retrieving candidate videos through an efficient video retrieval method. We employ established techniques for large-scale

search, such as IVFPQ index structures, at this stage of our method. Following this, our approach exhibits constant time $O(1)$ complexity relative to the number of videos in the collection. This is because our DRAQ-based re-ranking computation only applies to a fixed number of top-k candidate videos, thus not impacting scalability. Furthermore, the computational cost of DRAQ (and DTW) is negligible when compared to the computation required for frame-level feature extraction [2]. The computation time required for DRAQ is just 0.0546% of the feature extraction time. Finally, the overhead of DRAQ compared to DTW is negligible since it just requires the sampling of random paths through the already computed cost matrix C.

Regarding the storage impact of our method, we observe that only the aggregate features must be stored for candidate retrieval (similar to other video retrieval methods). The frame-level features of the top-k candidates (for DRAQ) can optionally be computed on the fly, trading off compute for storage cost.

C.2 Details of NMS [2]

NMS enhances a single-frame model by incorporating a temporal Transformer, trained through framewise temporal self-supervision. The primary reason for utilizing [2] in our method is its ability to deliver state-of-the-art video retrieval results and provide robust frame-wise features. This effectiveness largely stems from its capability to disrupt shortcuts in temporal pretext tasks, thereby promoting accurate frame-level temporal correspondence. For our task, we employ [2] equipped with a ViT-L backbone, pretrained on Kinetics400.

D Limitations

Our framework assumes that the top matches for alignable videos can be synchronized by identifying clips that are temporally alignable with the query. We presume that within these top matches, all frames are capable of aligning to the query and exhibit temporal monotonicity. However, challenges arise in real-world scenarios where videos may differ in the execution order of action classes. Additionally, some videos may contain extraneous segments that do not match the query (e.g., a video of “cutting pineapple" might include collecting the fruit from the fridge, which the query does not feature). Despite these complexities, we find that DRAQ is an effective tool for identifying alignable videos in simpler scenarios of non-partial and monotonic alignments. Leveraging a sufficiently expansive search space, this methodology has the potential to enhance the general applicability of video alignment techniques. We identify the resolution of partial matching and non-monotonic alignment as important directions for future extensions.

References

1. Chen, M., Wei, F., Li, C., Cai, D.: Frame-wise action representations for long videos via sequence contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13801–13810 (2022)
2. Dave, I.R., Jenni, S., Shah, M.: No more shortcuts: Realizing the potential of temporal self-supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1481–1491 (2024)
3. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)