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/

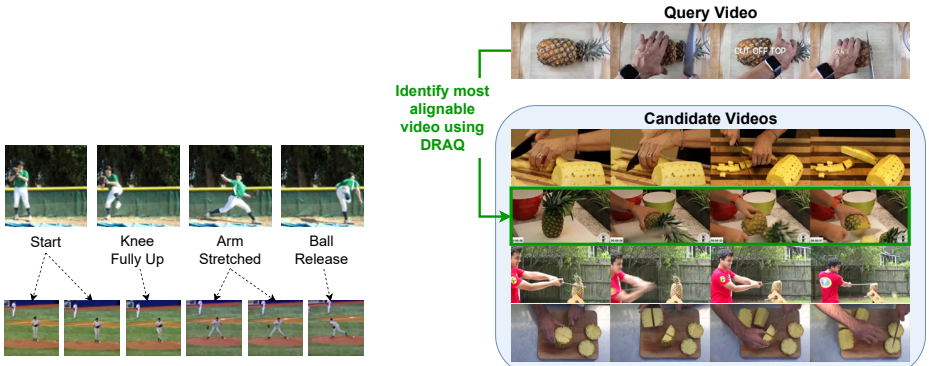
Abstract. Temporal video alignment aims to synchronize the key events like object interactions or action phase transitions in two videos. Such methods could benefit various video editing, processing, and understanding tasks. However, existing approaches operate under the restrictive assumption that a suitable video pair for alignment is given, significantly limiting their broader applicability. To address this, we re-pose temporal alignment as a search problem and introduce the task of Alignable Video Retrieval (AVR). Given a query video, our approach can identify well-alignable videos from a large collection of clips and temporally synchronize them to the query. To achieve this, we make three key contributions: 1) we introduce DRAQ, a video alignability indicator to identify and re-rank the best alignable video from a set of candidates; 2) we propose an effective and generalizable frame-level video feature design to improve the alignment performance of several off-the-shelf feature representations, and 3) we propose a novel benchmark and evaluation protocol for AVR using cycle-consistency metrics. Our experiments on 3 datasets, including large-scale Kinetics700, demonstrate the effectiveness of our approach in identifying alignable video pairs from diverse datasets.

Keywords: Temporal Alignment · Video Understanding

1 Introduction

Video understanding has made great progress in recent years, as evidenced by numerous tasks and benchmarks ranging from action recognition [30, 38] and localization [4, 20] to video editing [5] and generation [36]. A key challenge in the semantic and temporal understanding of videos is that of temporal video alignment. We say that two videos are temporally aligned when their key events (*e.g.*, action phase transitions, object interactions, etc.) co-occur exactly. For example, given two videos showing a "baseball swing," a video alignment approach would thus warp the videos so that the moment the person starts the swing and the moment the ball is released happen simultaneously in the two videos (see

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



(a) Regular Temporal Alignment: A pair of videos from the same action class is given. The goal is to align them, *i.e.*, match their key-event frames

(b) Proposed Alignable Video Retrieval (AVR): Given a query video, the goal is to find the best alignable video from candidate videos of the video search results.

Fig. 1: Alignable Video Retrieval. While some actions, like "baseball swing" (left), permit temporal alignment in virtually all cases due to their fixed sequence of action phases, general videos from other action classes, like "cutting pineapple" (right), exhibit much more variability. Knowledge of the action category alone is insufficient to identify alignable pairs for these cases, and a deeper temporal understanding of the videos is required to identify alignable videos. We propose DRAQ, an alignability score that can reliably identify the alignable video pair (red) among the set of candidates.

Figure 1a). Such methods can enable various applications in video processing and analysis. For example, it enables example-based video retiming [28] where a given video is warped according to dynamics found in another video, automated video clip replacement without the need for editing (*e.g.*, when license issues prevent the use of the original clip), the automatic time-aligned transfer of audio tracks [14] or video effects between clips, to name a few.

However, existing approaches for video alignment [6,14,23] primarily consider a restrictive, weakly supervised setting, where the pair of videos for alignment is assumed given, and only the alignment of the two videos is to be found. Because identifying alignable videos is a challenging problem that has remained unaddressed thus far, this limits the broader applicability of video alignment methods. Furthermore, existing methods train and test predominantly on a limited number of well-behaved action categories with well-delignated action phases. General real-world videos are often not as well-behaved as the types of videos considered in existing video alignment benchmarks. For example, they might not conform to a fixed sequence of key events and might exhibit large variations in how an action is performed. As illustrated in Figure 1, while videos of actions like "baseball swing" are almost always alignable due to their shared sequence of action phases, for videos of more general categories like "cutting pineapple," the larger variation in videos makes knowledge of the action category insufficient for identifying alignable video pairs. Given these observations, off-the-shelf video retrieval methods (*e.g.*, trained for action recognition) by themselves are insuffi-

cient for identifying alignable video pairs and a tailored solution for filtering or re-ranking candidate pairs is required. Therefore, in this paper, we address the question: *How to identify alignable videos from a large collection?*

Toward this goal, we introduce the task of Alignable Video Retrieval (AVR), tackling the thus far unaddressed problem of *identifying* alignable videos from a large dataset. The AVR task naturally combines the fundamental video understanding problems of retrieval and alignment, and we make several technical contributions toward a solution. As a first contribution, we propose a method for identifying the best alignable clips from a set of retrieved candidate videos. To this end, we introduce an alignability indicator that scores how alignable two video clips are. Our method, Dynamic Relative Alignment Quality (DRAQ), compares the optimal alignment cost (as obtained through Dynamic Time Warping [33]) to the average cost of multiple random sub-optimal alignments. Experimentally, we show that this intuitive indicator effectively identifies the best alignable videos by showing its high correlation with the action phase agreement of aligned videos and performance improvements in reranking video search results. As a second contribution, we propose a method to effectively improve the performance of off-the-shelf video-frame representations for video alignment. To this end, we introduce a feature contextualization approach, which augments a given frame-level feature representation with additional temporal context. Experimentally, we show that various off-the-shelf representations benefit from such contextualization for video alignment. Finally, as our third contribution, we propose a set of benchmarks and evaluation protocols to measure AVR performance. Our evaluation includes existing datasets with dense action phase labels where we propose aligned phase agreement to measure alignment quality and newly annotated Kinetics videos to evaluate the full AVR pipeline in a cycle-consist manner on a more diverse set of natural videos.

2 Prior Work

Our work on Alignable Video Retrieval (AVR) relates to several topics in computer vision, including video alignment, retrieval, and temporal feature learning.

Video Alignment. Several recent works proposed targeted training strategies to learn video representations for alignment. Many of these works study a weakly supervised setting, where video pairs showing the same action are used for learning [14, 16, 23, 43]. These methods rely on cycle consistency [14], DTW-based temporal alignment constraints [16, 23], or a combination of the two [21]. Another line of work focuses on purely unsupervised learning of frame-level features for alignment on unlabelled videos. These works build on variations of frame-contrastive learning objectives on augmented video clips [6, 11], some also incorporating weak supervision [43]. Alignment with the resulting frame-level features is typically performed via Dynamic Time Warping (DTW) [33]. Some methods also leverage differentiable formulations thereof [7], and variants robust to outliers have been proposed [13]. Instead of designing a novel learning strategy for video alignment, we build our approach on existing pre-trained video

representations, for which we introduce a video contextualization procedure to enhance their alignment capabilities.

Video Retrieval. A comprehensive AVR solution relies on good candidate proposals obtained through large-scale video-to-video retrieval. While early video search approaches were based on hand-crafted features [37], recent methods use neural-network representations, *e.g.*, learned through action recognition [9,17,41] or self-supervised learning [8, 25, 26, 34, 40]. Of particular interest are visual retrieval methods that also aim at localizing a matching segment in a video. Such approaches are prevalent in the video copy detection literature [1, 12, 13, 22, 24, 29, 39]. Most related are two-stage approaches [3, 39], where an initial coarse retrieval and localization are then refined with an "alignment" stage. Video copy detection approaches have in common that they are looking for identical video segments (up to video processing artifacts). Instead, our goal is to retrieve videos that permit a semantic, temporal alignment between different videos.

Temporal Video Representation Learning. A key requirement for accurate video alignment is a per-frame representation that captures the temporal features in a video and can discriminate subtle changes as a scene evolves over time. Several works explored the use of temporal self-supervision to learn such temporal sensitivity, *e.g.*, through pretext tasks about the ordering of video frames [18, 31, 32], or the classification of playback direction [42] and speed [2, 15, 27]. We leverage video features learned through such temporal self-supervision [10] both for retrieval and alignment in our approach.

3 Temporally Aligned Video Retrieval

We propose a solution for temporally Aligned Video Retrieval (AVR). Given a query video clip, our approach aims to identify the best alignable clip among a large collection. It then temporally aligns this best-alignable clip to the query, accurately transferring the timing of key events in the query clip. Our approach for retrieval and alignment of videos consists of three stages: 1) large-scale retrieval of candidate video clips from a large collection, 2) re-ranking of candidates based on a novel alignability indicator DRAQ, and 3) aligning of the best pair using DTW on contextualized frame-level features. An overview of our system is provided in Figure 2.

We build on video representations obtained with temporal self-supervised pre-training as described in [10] to represent whole video clips and their frames. These representations are shown in [10] to achieve state-of-the-art results in video action retrieval, and as our experiments in Section 5 show, they also achieve state-of-the-art in video alignment provided an additional global video contextualizing of the per-frame features we introduce. Given a video $V_i \in \mathbb{R}^{T \times H \times W \times 3}$ consisting of T frames of size $H \times W$, we encode it with the encoder E from [10] to obtain the feature sequence

$$F_i = [f_1^{(i)}, \ldots, f_T^{(i)}] \in \mathbb{R}^{T \times d}, \tag{1}$$

where d is the size of each per-frame feature vector $f_j^{(i)}$.

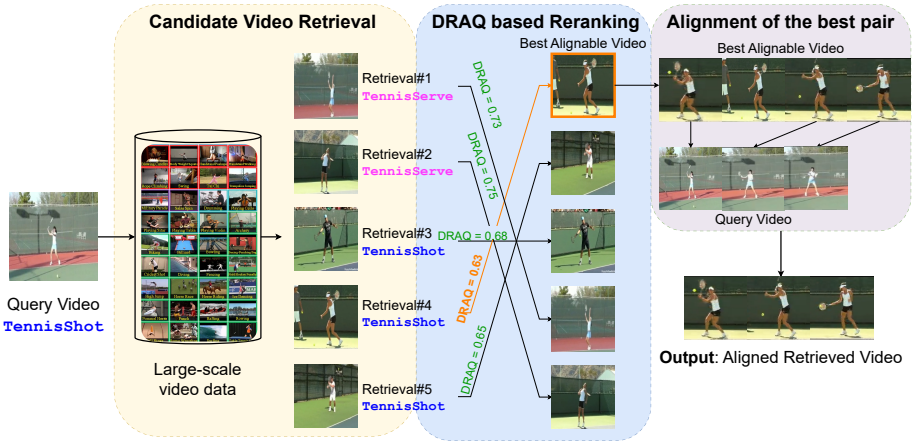


Fig. 2: Model Overview. We introduce a model for Aligned Video Retrieval (AVR): Given an input query video clip, our model aims to find and temporally align the best matching video among a large collection of videos. Our approach has three stages: 1) candidate retrieval from a large-scale database, 2) re-ranking of the top candidates to identify the most alignable clip using our procedure DRAQ, and 3) alignment of query and top match using DTW on our contextualized frame-level features.

Candidate Video Retrieval. To retrieve candidate video clips from a large collection of clips, we build a search index using temporally aggregated feature vectors $\bar{F}_i = \frac{1}{T}\sum_{j=1}^{T} f_j^{(i)}$. These clip-level feature vectors are stored in an efficient approximate nearest-neighbor data structure. As an additional preprocessing step, we standardize the feature vectors based on the mean and standard deviation computed on the retrieval dataset. Given a query video, we use cosine similarity to find the top-k candidates for alignment from the dataset.

Contextualized Frame-Level Features for Alignment. Given two videos, V_1 and V_2, with n and m frames, respectively, we now describe how to construct contextualized features for each frame that will be used for temporal alignment. We augment the base frame-level features $f_i^{(j)}$ from E (see Equation 1) with sequence-level context to better support the alignment task. Concretely, each feature should not just capture scene features at a specific moment (*e.g.*, the pose of a person at some point in an action sequence) but also how that moment fits into the overall action sequence (*e.g.*, is it at the end or beginning). To endow the video features with such temporal context, we concatenate them with the cumulative sum of features up to each time step. Concretely, our contextualized features for a video with T frames are given by

$$\bar{f}_j^{(i)} = f_j^{(i)} \oplus \frac{1}{T}\sum_{t=1}^{j} f_t^{(i)} \in \mathbb{R}^{T\times 2d}, \tag{2}$$

where $\oplus$ indicates concatenation along the channel dimension. Finally, we standardize the frame features per clip via zero-centering, *i.e.*, working with

$$\hat{f}_i^{(1)} = \bar{f}_i^{(1)} - \frac{1}{T}\sum_{l=1}^{T} \bar{f}_l^{(1)}, \tag{3}$$

instead of $f_i^{(1)}$ in the following ($\hat{f}_j^{(2)}$ is defined similarly). Note that this approach is very general and can be applied to any frame-level representation and video length, even when the sequence length at inference time is very different from pre-training (as is the case with our default video features).

Temporal Alignment via DTW. Given pairs of candidate video clips from our retrieval stage, we leverage Dynamic Time Warping (DTW) to find an optimal alignment between the two frame sequences. Since our alignability indicator (DRAQ) is closely related to the computations required for DTW, we provide a detailed description of DTW first. DTW operates on a cost matrix $C \in \mathbb{R}^{n\times m}$, which quantifies the similarity between feature vectors from the two videos. To compute C, we employ the frame-level distance

$$C(i,j) = 1 - \frac{\hat{f}_i^{(1)} \cdot \hat{f}_j^{(2)}}{\|\hat{f}_i^{(1)}\|\|\hat{f}_j^{(2)}\|}, \tag{4}$$

where $\hat{f}_i^{(1)} \cdot \hat{f}_j^{(2)}$ denotes the dot product between the two feature vectors, and $\|\cdot\|$ is the vector norm. DTW then determines the optimal alignment path P_{DTW}, which minimizes the cumulative cost through C from the top-left $(1,1)$ to the bottom right (n,m). More concretely, a path P through a cost matrix C of size $n \times m$ is defined as a sequence of tuples

$$P = ((i_1, j_1), (i_2, j_2), \ldots, (i_L, j_L)),$$

where:

- $i_1 \leq i_2 \leq \ldots \leq i_L$ and $j_1 \leq j_2 < \ldots \leq j_L$
- L is the length of the path with $L \leq n + m$.
- The start and end points are fixed: $(i_1, j_1) = (1,1)$ and $(i_L, j_L) = (n,m)$.

To compute the optimal path P_{DTW}, let D be a matrix of the same dimensions as C where each entry $D(i,j)$ represents the minimum cumulative cost of aligning the sequences up to $\hat{f}_i^{(1)}$ and $\hat{f}_j^{(2)}$. This is computed recursively as

$$D(i,j) = C(i,j) + \min_{(u,v)\in\Delta} D(i-u, j-v), \tag{5}$$

where $\Delta = \{(0,1), (1,0), (1,1)\}$ is a set of offsets corresponding to the three valid moves.

The optimal path P_{DTW} is then traced back from $D(n,m)$ to $D(1,1)$, choosing at each step the direction that resulted in the minimal cumulative cost. This path then also defines the optimal temporal alignment between the two frame

sequences in our method. In some cases, we might want to restrict ourselves to solutions that keep one of the involved frame sequences unwarped, *e.g.*, in example-based video re-timing or similar applications. We opt to skip any still frames for that particular sequence in the optimal path for such cases (*i.e.*, we delete such index tuples from the path).

DRAQ: Assessing Alignability for Re-Ranking. For our approach to AVR to work, it is crucial to have a way to assess the quality of an alignment between two videos. A straightforward choice would be to use the optimal DTW path cost $D(n, m)$. However, $D(n, m)$ builds on the absolute similarity of frames between the two sequences, which is more strongly influenced by the appearance of the two clips rather than their temporal alignability.

Therefore, we propose a new method to assess the quality of an alignment between two videos, Dynamic Relative Alignment Quality (DRAQ). The idea behind DRAQ is to compare the optimal alignment (as obtained with DTW) to an average random alignment between the two videos. Intuitively, if the optimal alignment achieves a clearly lower cost than a random alignment, we can be more confident that a meaningful synchronization of key video moments could be achieved. To determine this baseline cost of a random alignment, we generate k random alignment paths in the cost matrix C. To generate a random path, we start at $(i, j) = (n, m)$ and, at each step, select one of the possible moves $(\delta_i, \delta_j) \in \{-1, 0\}^2$ as follows:

- Sample $\delta_i = -1$ with probability $P_{\text{up}} = \frac{i}{i+j}$
- Sample $\delta_j = -1$ with probability $P_{\text{left}} = \frac{j}{i+j}$
- Move into direction (δ_i, δ_j) until $(1, 1)$ is reached

We ignore any steps equalling $(\delta_i, \delta_j) = (0, 0)$ in this process. Note how the paths are being biased towards the top-left direction by making the direction probabilities proportional to i and j. This is important to make the random paths more "challenging" compared to completely random sampling. To compute the cost of a randomly sampled path, we sum up all the corresponding entries in C. This process is repeated k times to generate k random path costs. The costs of the k random paths are finally averaged to obtain $\text{Cost}_{\text{random}}$.

The DRAQ metric is then defined as the ratio of the cumulative cost along the optimal alignment path to the average cost of k random paths, *i.e.*,

$$\text{DRAQ} = \frac{D(n, m)}{\text{Cost}_{\text{random}}}, \tag{6}$$

where $D(n, m)$ is the cumulative cost of the optimal alignment path. Because C is computed only once and sampling random paths through C is very efficient, DRAQ has minimal computational overhead compared to DTW.

With this formulation, the DRAQ score provides a robust mechanism to quantify video alignment quality by comparing the efficacy of a globally optimal alignment scheme (DTW) to average random alignments. Furthermore, because DRAQ is defined as a ratio of path costs, it does not suffer from the same appearance bias as DTW and serves as a better alignability indicator, as demonstrated

in our experiments.

The Aligned Video Retrieval Pipeline. To summarize, given a query video, we build an AVR pipeline consisting of 1) candidate alignable video retrieval using k-nearest neighbor retrieval on clip-level embeddings, 2) re-ranking and filtering of candidates for alignability using DRAQ on contextualized per-frame features, and 3) warping of the best match using DTW.

4 Evaluating Aligned Video Retrieval

We propose a protocol to evaluate aligned video retrieval methods. Existing video alignment benchmarks do not tackle the problem of *identifying* alignable video pairs and instead build on existing video datasets with well-behaved, alignable action classes like PennAction, where alignability for videos of the same action is assumed. Prior works typically report alignment performance via several proxy tasks involving the phase labels, such as phase classification via learned predictors or frame retrieval. While we also use PennAction and the associated temporal action-phase labels to evaluate our alignment component, we propose a more direct way to measure alignment quality via phase agreement of aligned videos. Furthermore, we introduce additional benchmarks for candidate retrieval reranking methods like our proposed alignability indicator, DRAQ.

Problems in Existing Alignment Benchmarks. Established video alignment protocols on PennAction primarily consider proxy metrics, such as action phase classification or Kendell Tau, for measuring alignment performance. We observe in our experiments that such metrics can to a large extent be gamed, provided sufficient knowledge of each frame's position in the video is encoded in the frame features. For example, we find that a BYOL [19] frame encoder combined with a temporal Transformer processing the frame embeddings (an architecture analogous to SotA methods like CARL [6]) achieves very high performance *even with a random initialization of the Transformer*. We suspect the position encoding's influence on the frame embeddings is the reason for this phenomenon. As a result, such proxy tasks might not provide a good indication of real-world alignment performance.

Instead, for datasets like PennAction with dense phase labels, we propose to evaluate the alignment directly by computing phase label agreement after alignment. Concretely, we take pairs of videos, compute their temporal alignment according to DTW on the extracted frame features, and report the average agreement of the phase labels after alignment. We term this metric Aligned Phase Agreement (APA).

Cycle Consistency for AVR Evaluation. Ideally, an AVR benchmark would also consist of videos with dense action phase annotations. However, obtaining high-quality phase annotations for large-scale and diverse videos is very costly. Furthermore, it is difficult to consistently define action phases or key events for general videos, as is required for such an approach. Instead, we propose to leverage cycle consistency as a proxy for AVR performance. Concretely, given a

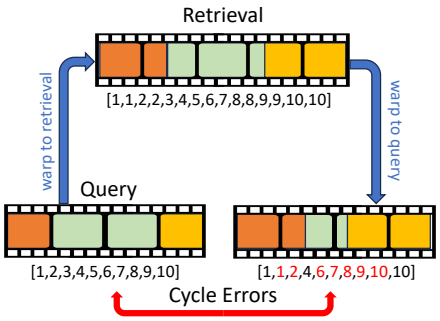


Fig. 3: AVR evaluation via Cycle-Consistency. We illustrate the use of consistency errors to measure aligned video retrieval performance. A query video (bottom left), along with phase labels (colored regions) and frame indices (below the video), is warped to the top retrieval video (top). The aligned labels and frame indices are then warped back to the query again to complete the cycle. We then report the Frame Position Error (FPE) and the Cycle Phase Error (CPE) when the query contains phase information.

query video, we 1) obtain the top candidate using the AVR model, 2) align the query to the top match and propagate per-frame labels (*e.g.*, position or phase labels) to the match, and 3) cycle back to the query with another alignment and label propagation. Note that we perform both warps in this cycle so that the second video is kept unwarped, thus ensuring that the cycle-warped video has the same length as the query. This is illustrated in Figure 3.

We propose to use this cycle consistency in two settings: 1) label propagation where the query contains phase annotations but the retrieved match does not, and 2) the position of each frame in the query is used as a "label" for propagation (no human labeling required). For scenario 1, we report the Cycle Phase Error (CPE) as the average error in phase labels before and after the cycle, and for scenario 2, we report Frame Position Error (FPE) as the MSE between the original and cycle-warped frame position vector. Since scenario 1 with CPE only requires the query to contain phase annotation, the approach easily scales to large retrieval datasets and avoids the need for consistent phase annotation between query and retrieval. We leverage these benefits to quantify the performance on more general natural videos by labeling a set of Kinetics validation videos with intuitive phase labels (*i.e.*, choosing characteristic key moments in the videos).

5 Experiments

We performed cycle-consistency experiments to demonstrate the effectiveness of our solution for Aligned Video Retrieval (AVR) in Section 5.2. Additionally, we report experiments verifying the two key components of our method: 1) frame-level contextualized video feature design (Eq. 3) for video alignment in Section 5.3, and 2) DRAQ for identifying video pairs with the highest alignment quality in Section 5.4.

5.1 Datasets

We consider three well-known video datasets in our experiments:

Candidates	Alignment Features	Reranking Metric	PennAction↻ FPE ↓	PennAction↻ CPE ↓	Penn ⇄ UCF FPE ↓	Penn ⇄ UCF CPE ↓	Kinetics↻ FPE ↓	Kinetics↻ CPE ↓
NMS [10]	BYOL [19]	-	0.5	0.40	121.1	105.0	13.0	1.03
		DRAQ	0.2	0.13	50.6	11.0	0.3	0.09
	CARL [6]	-	90.3	2.38	18.7	28.5	23.5	0.45
		DRAQ	24.3	0.74	5.2	5.9	2.3	0.08
	NMS [10]	-	13.4	1.32	5.5	22.2	22.7	0.86
		DRAQ	9.5	0.20	4.8	5.9	0.5	0.0
Oracle	BYOL [19]	-	50.4	4.14	–	–	7.6	0.62
		DRAQ	7.5	0.53	–	–	0.3	0.05
	CARL [6]	-	23.4	1.34	–	–	36.4	1.04
		DRAQ	11.2	0.36	–	–	1.7	0.14
	NMS [10]	-	24.7	1.70	–	–	35.3	1.08
		DRAQ	7.8	0.33	–	–	0.3	0.01

Table 1: Cycle Consistency for AVR Evaluation. We report cycle consistency errors for cycle-warped phase labels (CPE) and frame positions (FPE) on PennAction, Kinetics, and between PennAction and UCF101. The symbol ↻ means that the query video and retrieval set are from the same dataset, while ⇄ shows that the query video and retrieval set are from different data sources. We show results for AVR candidates obtained with retrieval using clip-level NMS features and oracle retrieval, which randomly chooses candidates that show the same action as the query. The performance of BYOL, CARL, and NMS features, all using our feature contextualization, is reported. For each case, we show the effectiveness of DRAQ reranking, which is applied to the $k = 10$ candidates to choose the top match.

UCF101 [38] is an action recognition dataset containing 13,320 videos of 101 human actions, which are collected from internet videos. We use split-1 for our experiments, where there are 9,537 training videos and 3,783 test videos.
PennAction [44] containing videos of various sports. We use the same split as prior video alignment work [6, 14], covering 2,106 videos with 13 action classes. Each video has associated video-level action and frame-level action-phase labels.
Kinetics700 [30], a large-scale dataset with about 650,000 natural videos of 700 diverse action classes from the internet. We annotate 91 validation videos with intuitive key-frames to define phase labels for cycle-consistent AVR evaluation.

5.2 Alignable Video Retrieval Evaluation via Cycle-Consistency

We evaluate our full AVR pipeline (Section 3) in cycle-consistency protocols as described in Section 4 on PennAction, UCF101, and Kinetics700. We report results in Table 1. For candidate proposals, we perform video retrieval (with clip-level features) as described in Section 3 using the state-of-the-art self-supervised representations of [10] (dubbed NMS). On PennAction and Kinetics, we also report results with an Oracle proposal generator, choosing candidates from the same class as the query (not reported for Penn⇄UCF since action classes differ). In all cases, query videos are taken from the validation set, and retrieval is performed on the training set of the respective datasets. We compare several off-

the-shelf frame-level features [6,10,19] in combination with our proposed DRAQ re-ranking (applied to the top-10 candidates) regarding temporal alignment performance. We apply our feature contextualization to all the features since we found it beneficial in all cases (see also Section 5.3). We report average Frame Position Error (FPE) and Cycle Phase Error (CPE) using the phase annotation of the query video.

We can observe clear improvements for DRAQ re-ranked candidates in all cases, which suggests that DRAQ successfully identified the videos that are better aligned among the candidate set. While we do observe a lot of variability between feature models and datasets, contextualized NMS features with DRAQ re-reranking appear to perform best overall.

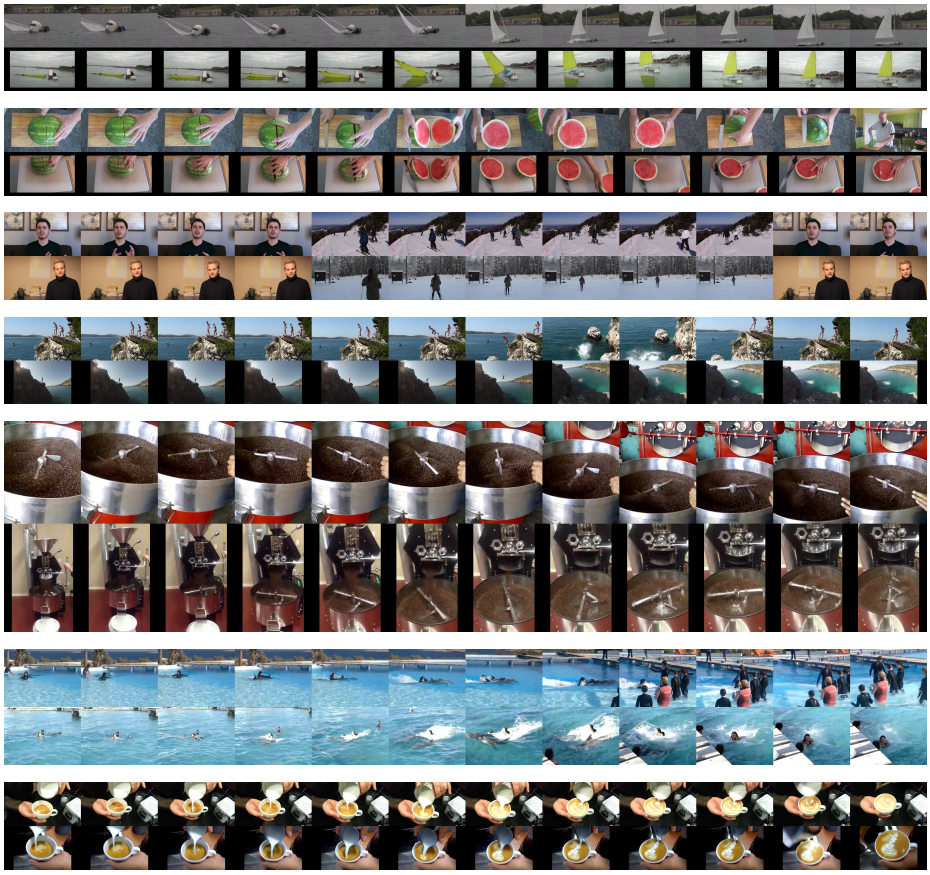


Fig. 4: Qualitative Examples of Aligned Video Retrieval on Kinetics700. The top frame sequence in each row shows the query video (from the validation split), and the bottom sequence shows the aligned retrieval (from the training split) with the lowest DRAQ score among the retrieved candidates. We show results for video pairs with DRAQ$<$ 0.6, which generally suggests meaningful alignment (zoom in for detail).

Features	+context	Avg.	Top-DRAQ
BYOL [19]	✗	0.769	0.814
CARL [6]	✗	0.826	0.856
NMS [10]	✗	0.832	0.868
BYOL [6]	✓	0.801	0.821
CARL [6]	✓	0.827	0.854
NMS [10]	✓	0.848	0.893

Table 2: Aligned Phase Agreement. We compare frame-level features with and without our contextualization. We report the Aligned Phase Agreement (APA), *i.e.*, the average agreement of phase labels after warping pairs of videos using DTW. Results are computed over the top-10 candidate pairs in our AVR setting (query videos are from PennAction-`val`, and retrieval is over `train`). We report average APA over all candidates and for the top DRAQ re-ranked example per query.

AVR appears to be notably difficult between PennAction and UCF. We hypothesize that this is due to the combination of action class mismatch and the limited size of the retrieval set. Qualitative AVR examples on Kinetics are provided in Figure 4 and in the Supplementary, where we also provide an expanded table including DTW re-ranking and non-contextualized features.

5.3 Contextualized Frame-Level Features for Video Alignment

To evaluate the improvements due to our proposed contextualized frame-level features, we perform experiments using the action phase annotations on PennAction. We report average Aligned Phase Agreement accuracy (APA) after DTW alignment for the top-10 retrieval candidate pairs in our AVR setting. In Table 2, we compare our contextualized features (+context) to the baseline performance of non-contextualized features from different frame-level feature methods. We compare features from NMS [10] trained through temporal self-supervision against BYOL [19] (a strong self-supervised image representation) and the recent state-of-the-art video alignment method CARL [6]. We can observe clear improvements with our contextualization (Equation 3) for NMS and BYOL, which benefit from the added temporal context. Since CARL already possesses temporal context due to a longer-range temporal Transformer, we do not observe additional benefits.

We also compare contextualized features with prior video alignment methods on various established proxy tasks in Table 3. As pointed out in Section 4, we observe shortcuts based on frame position information for models leveraging temporal Transformers (*e.g.*, CARL). This is exemplified by the result for BYOL+Transformer, which combines BYOL with a *randomly initialized* untrained temporal Transformer. This variant represents the initialization of CARL and already outperforms all other prior works (notice the saturated τ values indicating the shortcut). We also report results with CARL when applied in a "sliding window" fashion (CARL-SW), similar to how other methods process

Method	Labels	Phase Classification (Top-1 Acc.)	Frame Retrieval Ret@1	Kendall Tau τ
TCC [14]	Action	0.744	0.767	0.641
GTA [21]	Action	-	-	0.748
LAV [23]	Action	0.786	0.791	0.684
TCN [35]	None	0.681	0.778	0.542
SaL [32]	None	0.682	-	0.474
BYOL [19]	None	0.545	0.473	0.216
CARL [6]	None	0.931	0.922	0.985
NMS [10]	None	0.799	0.730	0.397
BYOL+Transformer	None	0.863	0.817	0.995
CARL-SW	None	0.845	0.830	0.686
Our Contextualized Frame Features				
BYOL + context	None	0.881↑62%	0.782↑65%	0.776↑259%
CARL-SW + context	None	0.889↑5%	0.845↑2%	0.648↓5%
NMS + context	None	0.918↑15%	0.882↑21%	0.825↑102%

Table 3: PennAction Benchmarks. We demonstrate the effect of our feature contextualization on various frame representations in proxy tasks on PennAction, comparing contextualized features with prior temporal alignment works. These tasks assess the ability to decode temporal information (e.g., action phase) from frame features.

the videos. For methods with a limited temporal context (and thus not affected by the position shortcut), we can observe benefits from our feature contextualization. Given the observed shortcuts in the proxy tasks of Table 3, we argue for directly evaluating alignment performance as in Table 2 instead.

5.4 DRAQ for Measuring Video Alignment Quality

To verify the ability of DRAQ to identify well-alignable videos among a set of candidates (*e.g.*, obtained through retrieval), we compute the Aligned Phase Agreement (APA) after alignment obtained at different thresholds for DRAQ and alternative alignment quality indicators. Videos of the same action class have the same phase labels (assigned to each frame); thus, a high APA indicates that the aligned videos mostly show the same action phase (APA=1 means perfect phase alignment). Videos of different actions exhibit an APA of zero.

We plot the average APA for video pairs that fall below a given threshold for DRAQ and other alignment indicators in Figure 5. We compare DRAQ against the optimal DTW cost and Kendall τ (we use $-\tau$ to be consistent with DRAQ and DTW where lower values are better). For a fair comparison, we plot the thresholds as percentiles of the respective indicator values. The set of video pairs is taken from PennAction and is balanced, *i.e.*, we use the same number of pairs with matching and non-matching action classes. As the figure shows, pairs with low DRAQ values more consistently achieve high APA. This indicates that DRAQ is clearly superior in identifying alignable videos than the alternatives.

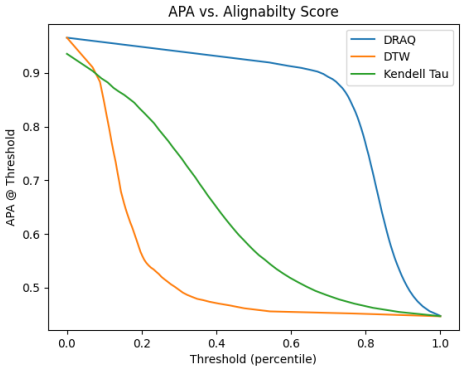


Fig. 5: DRAQ for Identifying Alignable Videos. We show a plot of the Aligned Phase Agreement (APA) for video pairs with alignment indicators below a given threshold on PennAction. The x-axis corresponds to the percentiles of the respective indicator. We compare DRAQ to the optimal DTW cost and Kendell Tau.

Method	**PennAction**		**UCF101**	
	R@1	R@10	R@1	R@10
w/o Reranking	82.09	98.76	81.60	93.52
DRAQ Reranking	82.40	99.17	81.81	93.92

Table 4: DRAQ for Action Retrieval Re-Ranking. We report recall@k with and without DRAQ re-ranking on PennAction and UCF101. DRAQ re-ranking is applied to the top 25 retrievals.

Effect of DRAQ on Action Retrieval. One of the applications of DRAQ, which is amenable to quantitative analysis, is re-ranking video search results. We show results on reranking the top-25 retrievals for PennAction and UCF101 in Table 4. We measure recall@k, where a retrieval is considered correct if the query and retrieval video shows the same action category. We observe improved recall in the top retrievals for both datasets with DRAQ re-ranking.

6 Conclusion

This paper explored the novel task of Aligned Video Retrieval (AVR) to tackle the problem of identifying temporally alignable videos from large collections. As a first step towards a solution, we introduced a video alignment score DRAQ, which, given a query video, can help us identify the most alignable videos among a set of candidates. In new cycle-consistency benchmarks to measure the performance of AVR, we show that DRAQ, together with carefully designed frame-level features, is able to identify alignable video pairs for general videos with diverse actions. With future work, we aim to advance AVR via improved candidate proposals from more sophisticated retrieval. Our work also holds particular interest in retrieval-augmented generation within diffusion models, where from an aligned retrieval video, one can effectively generate corresponding modalities, like audio, for the query video. We believe that further progress on this task will open up many novel applications for video alignment methods in video editing, discovery, and understanding.

References

1. Baraldi, L., Douze, M., Cucchiara, R., Jégou, H.: Lamv: Learning to align and match videos with kernelized temporal layers. In: Proc. CVPR. pp. 7804–7813 (2018)
2. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931 (2020)
3. Black, A., Jenni, S., Bui, T., Tanjim, M.M., Petrangeli, S., Sinha, R., Swaminathan, V., Collomosse, J.: Vader: Video alignment differencing and retrieval. arXiv preprint arXiv:2303.13193 (2023)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
5. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
6. 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)
7. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: International conference on machine learning. pp. 894–903. PMLR (2017)
8. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. Computer Vision and Image Understanding p. 103406 (2022). https://doi.org/https://doi.org/10.1016/j.cviu.2022.103406
9. Dave, I., Scheffer, Z., Kumar, A., Shiraz, S., Rawat, Y.S., Shah, M.: Gabriellav2: Towards better generalization in surveillance videos for action detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 122–132 (2022)
10. 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)
11. Dave, I.R., Rizve, M.N., Shah, M.: Finepseudo: Improving pseudo-labelling through temporal-alignablity for semi-supervised fine-grained action recognition. In: European Conference on Computer Vision (2024)
12. Douze, M., Revaud, J., Verbeek, J.J., Jégou, H., Schmid, C.: Circulant temporal encoding for video retrieval and temporal alignment. IJCV **119**, 291–306 (2015)
13. Dvornik, M., Hadji, I., Derpanis, K.G., Garg, A., Jepson, A.: Drop-dtw: Aligning common signal between sequences while dropping outliers. NeurIPS **34**, 13782–13793 (2021)
14. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1801–1810 (2019)
15. Epstein, D., Chen, B., Vondrick, C.: Oops! predicting unintentional action in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 919–929 (2020)
16. Fakhfour, N., ShahverdiKondori, M., Mohammadzade, H.: Video alignment using unsupervised learning of local and global features. arXiv preprint arXiv:2304.06841 (2023)

17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
18. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 5729–5738. IEEE (2017)
19. 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)
20. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6047–6056 (2018)
21. Hadji, I., Derpanis, K.G., Jepson, A.D.: Representation learning via global temporal alignment and cycle-consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11068–11077 (2021)
22. Han, Z., He, X., Tang, M., Lv, Y.: Video similarity and alignment learning on partial video copy detection. In: Proc. ACM Int. Conf. Multimedia. pp. 4165–4173 (2021)
23. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, Z., Tran, Q.H.: Learning by aligning videos in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5548–5558 (2021)
24. He, S., He, Y., Lu, M., Jiang, C., Yang, X., Qian, F., Zhang, X., Yang, L., Zhang, J.: Transvcl: Attention-enhanced video copy localization network with flexible supervision. In: AAAI. vol. 37, pp. 799–807 (2023)
25. Jenni, S., Black, A., Collomosse, J.: Audio-visual contrastive learning with temporal self-supervision. arXiv preprint arXiv:2302.07702 (2023)
26. Jenni, S., Jin, H.: Time-equivariant contrastive video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9970–9980 (2021)
27. Jenni, S., Meishvili, G., Favaro, P.: Video representation learning by recognizing temporal transformations. In: The European Conference on Computer Vision (2020)
28. Jenni, S., Woodson, M., Heilbron, F.C.: Video-retime: Learning temporally varying speediness for time remapping. arXiv preprint arXiv:2205.05609 (2022)
29. Jiang, Y.G., Wang, J.: Partial copy detection in videos: A benchmark and an evaluation of popular methods. IEEE Trans. Big Data **2**(1), 32–42 (2016)
30. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
31. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676 (2017)
32. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 527–544. Springer (2016)
33. Müller, M.: Dynamic time warping. Information retrieval for music and motion pp. 69–84 (2007)

34. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)
35. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 1134–1141. IEEE (2018)
36. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
37. Sivic, J., Zisserman, A.: Video google: Efficient visual search of videos. Toward category-level object recognition pp. 127–144 (2006)
38. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
39. Tan, W., Guo, H., Liu, R.: A fast partial video copy detection using knn and global feature database. In: Proc. WCACV. pp. 2191–2199 (2022)
40. Thoker, F.M., Doughty, H., Snoek, C.G.: Tubelet-contrastive self-supervision for video-efficient generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13812–13823 (2023)
41. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
42. Wei, D., Lim, J., Zisserman, A., Freeman, W.T.: Learning and using the arrow of time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8052–8060 (2018)
43. Zhang, H., Liu, D., Zheng, Q., Su, B.: Modeling video as stochastic processes for fine-grained video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2225–2234 (2023)
44. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE international conference on computer vision. pp. 2248–2255 (2013)