Video-Text Representation Learning via Differentiable Weak Temporal Alignment

Dohwan Ko[1] Joonmyung Choi[1] Juyeon Ko[1] Shinyeong Noh[1]
Kyoung-Woon On[2] Eun-Sol Kim[3] Hyunwoo J. Kim[1]*

[1]Department of Computer Science and Engineering, Korea University
[2]Kakao Brain [3] Department of Computer Science, Hanyang University
{ikodoh, pizard, juyon98, dneirfi, hyunwoojkim}@korea.ac.kr
{kloud.ohn}@kakaobrain.com {eunsolkim}@hanyang.ac.kr

Abstract

*Learning generic joint representations for video and text by a supervised method requires a prohibitively substantial amount of manually annotated video datasets. As a practical alternative, a large-scale but uncurated and narrated video dataset, HowTo100M, has recently been introduced. But it is still challenging to learn joint embeddings of video and text in a self-supervised manner, due to its ambiguity and non-sequential alignment. In this paper, we propose a novel multi-modal self-supervised framework **V**ideo-**T**ext **T**emporally **W**eak Al**i**gnme**n**t-based Contra**s**tive Learning (VT-TWINS) to capture significant information from noisy and weakly correlated data using a variant of Dynamic Time Warping (DTW). We observe that the standard DTW inherently cannot handle weakly correlated data and only considers the globally optimal alignment path. To address these problems, we develop a differentiable DTW which also reflects local information with weak temporal alignment. Moreover, our proposed model applies a contrastive learning scheme to learn feature representations on weakly correlated data. Our extensive experiments demonstrate that VT-TWINS attains significant improvements in multi-modal representation learning and outperforms various challenging downstream tasks. Code is available at https://github.com/mlvlab/VT-TWINS.*

1. Introduction

Learning video-text representations is an important problem in computer vision. In recent years, it has recently drawn increasing attention due to a large amount of video data and various applications. Previous works [32, 52, 57] have achieved exciting results by learning mappings between video clips and texts but they usually require a large amount of manual annotations such as MSR-VTT [55], DiDeMo [3], EPIC-KITCHENS [13]. However, since labeling videos is expensive and time-consuming, it does not scale well for sufficiently large datasets which are essential to learning generic video-text representations that are readily applicable to a wide range of downstream tasks that include text-to-video retrieval or video-text retrieval [27,50,51,56], text-based action localization [3,11], action segmentation [29, 43] and video question answering [34, 46, 56]. Recent studies suggest that multi-modal self-supervised learning with a huge amount of data is a promising alternative to fully supervised methods [15,54]. To this extent, HowTo100M [36] has been introduced, which is composed of 100 million pairs of video clips and captions from 1.22M *narrated instructional* videos.

The HowTo100M is one of the largest video datasets but it comes with several challenges. It is uncurated and its video-text pairs are weakly correlated meaning that given a video clip the caption depicting the visual content may appear *before/after* the clip or *not* even exist (Figure 1). To handle the weakly correlated video-text pairs, MIL-NCE [35] has proposed a multiple instance learning (MIL)-based contrastive learning adopting Noise Contrastive Learning (NCE) loss [19]. MIL-NCE treats the multiple captions which are temporally close to one clip as positive samples allowing one-to-many correspondence. But this strong assumption often leads to suboptimal representation learning.

In this paper, to address the problem, we develop a new weak temporal alignment algorithm building upon Dynamic Time Warping (DTW) [41]. In contrast to the standard DTW which is limited to sequential alignment, our proposed alignment algorithm allows flexibility by skipping irrelevant pairs and starting/ending at arbitrary time points. Also, it takes into account a globally optimal path as well as locally optimal paths by introducing local neighborhood smoothing. More importantly, our alignment algorithm is differentiable so we incorporate it into represen-

*is the corresponding author.

tation learning as a distance measure. We then propose a novel multi-modal self-supervised learning framework to learn a joint video and text embedding model named as **V**ideo-**T**ext **T**emporally **W**eak Al**i**gnme**n**t-based Contra**s**tive Learning (VT-TWINS) that automatically handles the correspondence between noisy and weakly correlated captions and clips.

Our extensive experiments on five benchmark datasets demonstrate that our learned video and text representations generalize well on various downstream tasks including action recognition, text-to-video retrieval, and action step localization. Moreover, ablation studies and qualitative analysis show that our framework effectively aligns the noisy and weakly correlated multi-modal time-series data.

Our **contributions** are threefold:

- We propose a novel self-supervised learning framework with differentiable weak temporal alignment that automatically handles the noisy and weakly correlated multi-modal time-series data.
- We analyze the local neighborhood smoothing in our alignment algorithm showing that unlike DTW the alignment takes into account local optimal paths as well as global optimal path.
- Our experiments show that the proposed method considerably improves joint representations of video and text an is adapted well on various downstream tasks.

2. Related Work

Self-Supervised Learning for Videos. The self-supervised learning approaches have received considerable attention because they do not require additional annotations during learning representation. Recently, several works are proposed to learn video representations in a self-supervised manner. One research direction is to design video-specific pretext tasks, such as verifying temporal orders [15, 30, 37, 54], predicting video rotation [24], solving jigsaw puzzles in a video [26], and dense predictive coding [21]. Another line of research is to use a contrastive learning which leads clips from the same video to be pulled together while clips from different videos to be pushed away [9, 10, 18, 23, 40, 44, 49]. In view of the multi-modality of videos, many works explore mutual supervision across modalities to learn representations of each modality. For example, they regard temporal or semantic consistency between videos and audios [8,28] or narrations [1,4,35,36] as a natural source of supervision. MIL-NCE [35] introduced contrastive learning to learn joint embeddings between clips and captions of unlabeled and uncurated narrated videos. The other line of work adopts an additional crossmodal encoder (*e.g*., crossmodal transformer) to capture richer interaction between modalities [17, 31, 33, 44, 45, 58]. In this paper, we focus on extending contrastive learning to temporally align two time-series modalities, *i.e*., clips and captions from videos without any additional crossmodal encoders.

Sequence Alignment. Sequence alignment is crucial in fields related to the time-series data due to the temporal information. In particular, the lack of manually annotated video datasets makes it harder to align clips and captions temporally. Dynamic Time Warping (DTW) [41] measures the distance with strong temporal constraints between two sequences. [7] uses global sequence alignment as a proxy task by relying on the DTW. [12, 20] extended the DTW for end-to-end learning with differentiable approximations of the discrete operations (*e.g*., the 'min' operator) in the DTW. Chang *et al*. [6] proposed the frame-wise alignment loss using the DTW in weakly supervised action alignment in videos. Drop-DTW [14] proposed a variant of the DTW algorithm which automatically drops the outlier elements from the pairwise distance to handle the noisy data. However, using the DTW alone can cause feature collapsing which leads all the feature embeddings to be concentrated to a single point. To address this problem, [6] and [22] use the subsidiary regularization loss term with the DTW.

3. Preliminaries

We briefly summarize the basic concepts of dynamic time warping and the characteristics of an uncurated narrated video dataset HowTo100M.

3.1. Dynamic Time Warping (DTW)

DTW [5] finds an optimal alignment between two time-series data. Let X and Y denote two time-series data of length n and m, *i.e*., $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_m]$. DTW first computes a pairwise distance matrix $\Delta(X,Y) := [\delta(x_i, y_j)]_{ij} \in \mathbb{R}^{n\times m}$ with a distance measure δ. Then, DTW optimizes the following:

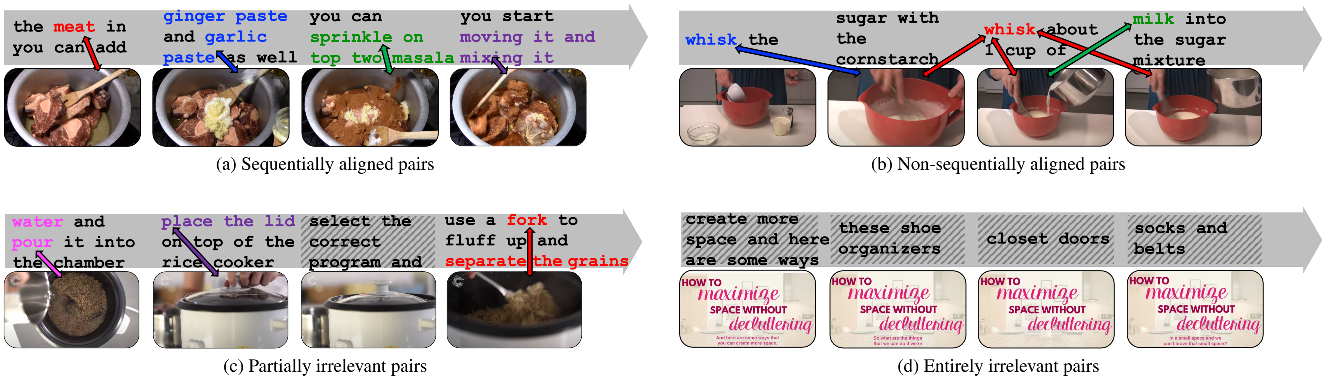
$$\text{DTW}(X,Y) = \min_{A\in\mathcal{A}_{n,m}} \langle A, \Delta(X,Y)\rangle, \tag{1}$$

where $\mathcal{A}_{n,m} \subset \{0,1\}^{n\times m}$ is a set of (binary) alignment matrices. An alignment matrix A represents a path that connects from $(1,1)$ to (n,m)-th entries of $\Delta(X,Y)$ by three possible moves $\{\downarrow, \searrow, \rightarrow\}$.

To efficiently find an optimal path, DTW [5] uses dynamic programming to recursively solve the following subproblems:

$$r_{i,j} = \delta_{i,j} + \min\{r_{i-1,j}, r_{i,j-1}, r_{i-1,j-1}\}, \tag{2}$$

where $r_{i,j}$ is the (i,j)-th element of a cumulative cost matrix $R(X,Y) \in \mathbb{R}^{n\times m}$ of $\Delta(X,Y)$. Therefore, $\text{DTW}(X,Y)$ in (1) is equal to $r_{n,m}$ which is the accumulated cost that evaluates the similarity between two time-series data.


the meat in you can add
ginger paste and garlic paste as well
you can sprinkle on top two masala
you start moving it and mixing it
whisk the
sugar with the cornstarch
whisk about 1 cup of
milk into the sugar mixture
water and pour it into the chamber
place the lid on top of the rice cooker
select the correct program and
use a fork to fluff up and separate the grains
create more space and here are some ways
these shoe organizers
closet doors
socks and belts
HOW TO maximize SPACE WITHOUT decluttering

(a) Sequentially aligned pairs

(b) Non-sequentially aligned pairs

(c) Partially irrelevant pairs

(d) Entirely irrelevant pairs

Figure 1. **Examples of the HowTo100M.** The HowTo100M dataset contains narrated instructional videos and the clips and captions are weakly aligned. Each video is composed of several pairs of clips and captions. Shaded captions are irrelevant to any clips in the video.

Soft-DTW [12] has proposed a *differentiable* variant of the DTW replacing the non-differentiable operator '$\min$' in (2) with the soft-min '$\min^\gamma$' defined as:

$$\min^\gamma\{a_1, a_2, \ldots, a_m\} = -\gamma \log \sum_{i=1}^{m} e^{-a_i/\gamma}, \qquad (3)$$

where $\gamma \in \mathbb{R}_+$ is a smoothing parameter. Then, the recurrence relation of Soft-DTW is given as:

$$r_{i,j} = \delta_{i,j} + \min^\gamma\{r_{i-1,j}, r_{i,j-1}, r_{i-1,j-1}\}. \qquad (4)$$

If γ is zero, soft-min $\min^\gamma$ is identical to $\min$ operator. As γ increases, Soft-DTW(X,Y) more takes into account the cost of suboptimal paths.

3.2. The HowTo100M Dataset

HowTo100M dataset [36] is a large-scale dataset that contains 136M video clips with paired captions from 1.22M narrated instructional videos across 23K different visual tasks. A video has 110 clip-caption pairs with an average duration of 4 seconds. The captions are automatically transcribed narrations via automatic speech recognition (ASR). Learning joint video text embeddings with HowTo100M has two sources of difficulties: 'uncurated narrations' and 'weak correlation' between clip-caption pairs. As discussed in [35], the narrations transcribed by ASR are potentially erroneous and the colloquial language is neither complete nor grammatically correct sentences. In addition, due to the weak correlation between the paired clips and captions, computing the optimal correspondence to learn joint embedding entails addressing the following challenges, which is the main focus of this paper.

Ambiguity. As aforementioned, the average duration of a clip-caption pair is 4 seconds. Since short clips are sampled densely in one video, consecutive clips are often semantically similar, *i.e.*, clip-caption alignments inherently have ambiguity. So it is more beneficial to use algorithms that take into account multiple alignments allowing many-to-many correspondence rather than the algorithms that consider the only one optimal path such as the standard DTW.

Irrelevant pairs. The paired clips and captions may contain irrelevant contents due to several reasons. People might *skip* to demonstrate some steps when narrations are clear enough or vice versa. In Figure 1c, since the narration "select the correct program" is clear enough, no demonstration is given in the corresponding clip. In addition, some videos have entirely irrelevant clips and captions like Figure 1d. When learning joint video text embeddings, these irrelevant pairs should be properly handled.

Non-sequential alignment. Although videos and texts are overall correlated at the video-level, the paired clips and captions often are not temporally well-aligned. For instance, people in a video describe plans *before* demonstrations or explain details *after* actions, *i.e.*, captions may come with temporal shifts. To estimate the correspondence between clips and captions, they can be aligned without changing the order of elements in each modality like Figure 1a, called *sequential* alignment. In contrast, when the order of elements in a modality is partially reversed or the content of a clip/caption is arbitrarily interspersed in the other modality, *non-sequential* alignments are required to compute the optimal correspondence. We observe that the non-sequential alignments often occur when videos have long sequences of captions and clips like Figure 1b. We will address the challenges by a new learning strategy.

4. Method

In this section, we present a novel multi-modal self-supervised framework, named as **V**ideo-**T**ext **T**emporally **W**eak Alig**n**ment-based Contra**s**tive Learning (VT-TWINS), to learn joint embeddings of video and text from uncurated narrated videos. To address the problems mentioned above and estimate more accurate correspondence, we propose a new differentiable variant of

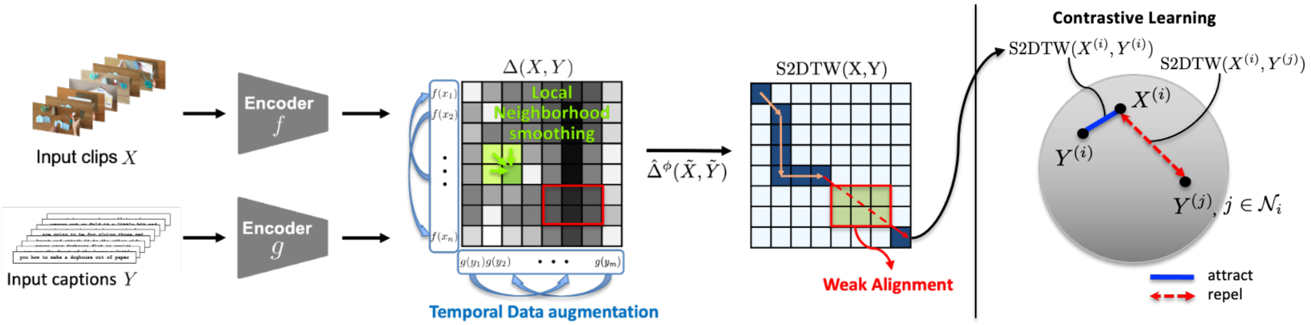

Input clips X
Input captions Y
Encoder f
Encoder g
$\Delta(X,Y)$
Local Neighborhood smoothing
$f(x_1)$
$f(x_2)$
$f(x_n)$
$g(y_1)\,g(y_2)$
$g(y_m)$
Temporal Data augmentation
$\hat{\Delta}^{\phi}(\tilde{X},\tilde{Y})$
S2DTW(X,Y)
Weak Alignment
Contrastive Learning
$\mathrm{S2DTW}(X^{(i)},Y^{(i)})$
$\mathrm{S2DTW}(X^{(i)},Y^{(j)})$
$X^{(i)}$
$Y^{(i)}$
$Y^{(j)}, j \in \mathcal{N}_i$
attract
repel

Figure 2. **Overall Architecture.** We propose a multi-modal self-supervised learning framework, VT-TWINS, to learn joint embeddings of video and text from noisy and weakly correlated data. The encoders f and g firstly extract feature embeddings from input clips X and input captions Y, respectively. Then, we present a new alignment algorithm based on the DTW, called S2DTW, which can handle weakly correlated data with local neighborhood smoothing (Section 4.1) and weak alignment (Section 4.2). We also apply temporal data augmentation (Section 4.3) to learn from the non-sequentially aligned data with the S2DTW. We finally employ a contrastive learning scheme, which uses the S2DTW as a distance measure between the clip-caption pairs, with negative pairs ($\mathcal{N}_i$) for representation learning while preventing feature collapsing (Section 4.4).

DTW, called Locally **S**moothed **S**oft-**DTW** with **W**eak Alignment (S2DTW). First, we apply local neighborhood smoothing and weak alignment. We then adopt temporal data augmentation for non-sequential alignments that the standard DTW cannot inherently handle. We finally apply a contrastive learning scheme and present VT-TWINS for representation learning without feature collapsing. Figure 2 and Algorithm 1 show our overall algorithm VT-TWINS including S2DTW.

4.1. Local Neighborhood Smoothing

To address the ambiguity as mentioned in Section 3.2, we smooth the pairwise distance matrix $\Delta(X,Y)$ as:

$$\hat{\delta}_{i,j} = \delta_{i,j} + \min{}^{\gamma}\{\delta_{i-1,j}, \delta_{i,j-1}, \delta_{i-1,j-1}\}, \quad (5)$$

where $\delta_{i,j}$ and $\hat{\delta}_{i,j}$ are the (i,j)-th elements of $\Delta(X,Y)$ and $\hat{\Delta}(X,Y)$, respectively. This allows many-to-many correspondence and encourages the alignment algorithm to focus more on a locally optimal clip (or caption), which has relatively smaller distances to others within a small neighborhood. $\hat{\delta}_{i,j}$ can be viewed as smoothed $\delta_{i,j}$ with its previous elements $\delta_{i-1,j}$, $\delta_{i,j-1}$, and $\delta_{i-1,j-1}$. Then, similar to (4) we apply dynamic programming to compute the optimal cost from smoothed distance matrix $\hat{\Delta}(X,Y)$ instead of $\Delta(X,Y)$ and as follows:

$$\hat{r}_{i,j} = \hat{\delta}_{i,j} + \min{}^{\gamma}\{\hat{r}_{i-1,j}, \hat{r}_{i,j-1}, \hat{r}_{i-1,j-1}\}. \quad (6)$$

S2DTW decays the cost of older matches and reflects more recent elements since (6) accumulates the cost from the top-left element to the bottom-right element, sequentially. Roughly speaking, the proposed S2DTW with $\hat{\Delta}(X,Y)$ considers local optimality by (5) as well as global optimality by (6) since S2DTW can be rewritten as:

$$\begin{aligned}\hat{r}_{i,j} = \delta_{i,j} &+ \min{}^{\gamma}\{\hat{r}_{i-1,j}, \hat{r}_{i,j-1}, \hat{r}_{i-1,j-1}\} \\ &+ \min{}^{\gamma}\{\delta_{i-1,j}, \delta_{i,j-1}, \delta_{i-1,j-1}\}.\end{aligned} \quad (7)$$

Differentiation. We compare Soft-DTW [12] and S2DTW via their derivatives. At the Soft-DTW, they denote a gradient matrix $M = [\mu_{i,j}]$ where $\mu_{i,j} := \frac{\partial r_{n,m}}{\partial \delta_{i,j}} = \frac{\partial r_{n,m}}{\partial r_{i,j}} \cdot \frac{\partial r_{i,j}}{\partial \delta_{i,j}} = \frac{\partial r_{n,m}}{\partial r_{i,j}} \cdot 1 = \frac{\partial r_{n,m}}{\partial r_{i,j}}$ by differentiating (4) w.r.t $\delta_{i,j}$. In S2DTW case, however, $\frac{\partial \hat{r}_{n,m}}{\partial \delta_{i,j}} \neq \frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i,j}}$ due to the local neighborhood smoothing layer, *i.e.*, $\frac{\partial \hat{r}_{i,j}}{\partial \delta_{i,j}} \neq 1$. We therefore redefine $\mu_{i,j} := \frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i,j}}$ and denote additional $\hat{\mu}_{i,j} := \frac{\partial \hat{r}_{n,m}}{\partial \delta_{i,j}}$ for the gradient matrix for local neighborhood smoothing layer. $\mu_{i,j}$ of S2DTW is calculated as follows:

$$\underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i,j}}}_{\mu_{i,j}} = \underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i+1,j}}}_{\mu_{i+1,j}} \cdot \frac{\partial \hat{r}_{i+1,j}}{\partial \hat{r}_{i,j}} + \underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i,j+1}}}_{\mu_{i,j+1}} \cdot \frac{\partial \hat{r}_{i,j+1}}{\partial \hat{r}_{i,j}} + \underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i+1,j+1}}}_{\mu_{i+1,j+1}} \cdot \frac{\partial \hat{r}_{i+1,j+1}}{\partial \hat{r}_{i,j}}. \quad (8)$$

By differentiating (6) with $i+1$ instead of i, the green term of (8) is calculated as:

$$\gamma \log \frac{\partial \hat{r}_{i+1,j}}{\partial \hat{r}_{i,j}} = \min{}^{\gamma}\{\hat{r}_{i,j-1}, \hat{r}_{i,j}, \hat{r}_{i+1,j-1}\} - \hat{r}_{i,j}. \quad (9)$$

After calculating $\mu_{i,j}$ in (8), $\hat{\mu}_{i,j}$ is calculated as:

$$\underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \delta_{i,j}}}_{\hat{\mu}_{i,j}} = \underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i+1,j}}}_{\mu_{i+1,j}} \cdot \frac{\partial \hat{r}_{i+1,j}}{\partial \delta_{i,j}} + \underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i,j+1}}}_{\mu_{i,j+1}} \cdot \frac{\partial \hat{r}_{i,j+1}}{\partial \delta_{i,j}} + \underbrace{\frac{\partial \hat{r}_{n,m}}{\partial \hat{r}_{i+1,j+1}}}_{\mu_{i+1,j+1}} \cdot \frac{\partial \hat{r}_{i+1,j+1}}{\partial \delta_{i,j}}. \quad (10)$$

In (10), $\frac{\partial \hat{r}_{i+1,j}}{\partial \delta_{i,j}} = \frac{\partial \hat{r}_{i+1,j}}{\partial \hat{\delta}_{i+1,j}} \cdot \frac{\partial \hat{\delta}_{i+1,j}}{\partial \delta_{i,j}} = \frac{\partial \hat{\delta}_{i+1,j}}{\partial \delta_{i,j}}$ since $\frac{\partial \hat{r}_{i+1,j}}{\partial \hat{\delta}_{i+1,j}} = 1$. Similar to (9), it is written as:

$$\gamma \log \frac{\partial \hat{r}_{i+1,j}}{\partial \delta_{i,j}} = \min{}^{\gamma}\{\delta_{i,j-1}, \delta_{i,j}, \delta_{i+1,j-1}\} - \delta_{i,j}, \quad (11)$$

Algorithm 1 VT-TWINS Algorithm with S2DTW

```
Inputs: clips $X$, captions $Y$
Parameters: smoothing parameter $\gamma$, dummy elements $\phi$

    # Temporal Data Augmentation
    $\forall i, \forall j \in$ mini-batch
    $\tilde{X}^{(i)}, \tilde{Y}^{(j)} \leftarrow Aug(X^{(i)}), Aug(Y^{(j)})$
    # Apply Contrastive Learning Scheme
    $\mathcal{L} \leftarrow -\log \sum_i \left( \frac{e^{-\mathrm{S2DTW}\left(\tilde{X}^{(i)}, \tilde{Y}^{(i)}\right)}}{e^{-\mathrm{S2DTW}\left(\tilde{X}^{(i)}, \tilde{Y}^{(i)}\right)} + \sum_{j \in \mathcal{N}_i} e^{-\mathrm{S2DTW}\left(\tilde{X}^{(i)}, \tilde{Y}^{(j)}\right)}} \right)$
    # S2DTW
    function S2DTW($X, Y$)
        $\delta_{i,j} \leftarrow \Delta(X,Y)[i,j], \ \forall i \in [1,n], \forall j \in [1,m]$
        # Local Neighborhood Smoothing
        for $(i,j) = (1,1)$ to $(n,m)$ do
            $\hat{\delta}_{i,j} \leftarrow \delta_{i,j} + \min^{\gamma}\{\delta_{i-1,j}, \delta_{i,j-1}, \delta_{i-1,j-1}\}$
        end for
        # Weak Alignment
        $\hat{\Delta}^{\phi} \leftarrow$ merge ($\hat{\Delta}$, $\phi$ )
        # Calculate DTW
        for $(i,j) = (1,1)$ to $(n,m)$ do
            $\hat{r}_{i,j} \leftarrow \hat{\delta}_{i,j} + \min^{\gamma}\{\hat{r}_{i-1,j}, \hat{r}_{i,j-1}, \hat{r}_{i-1,j-1}\}$
        end for
        return $\hat{r}_{n,m}$
    end function
Output: $\mathcal{L}$
```

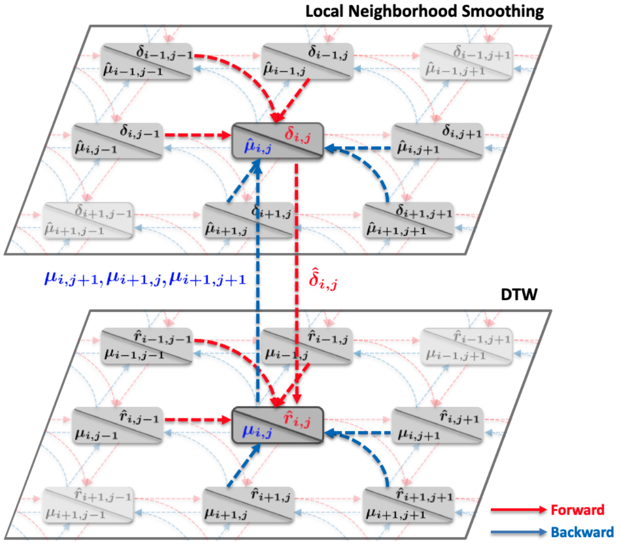


Figure 3. **Forward and Backward of Local Neighborhood Smoothing in terms of** (i,j)**.** At forward propagation, we firstly compute $\hat{\delta}_{i,j}$ by smoothing $\delta_{i,j}$ with $\delta_{i-1,j}$, $\delta_{i,j-1}$, and $\delta_{i-1,j-1}$ in (5). Then, $\hat{r}_{i,j}$ is calculated with smoothed $\hat{\delta}_{i,j}$ and $\hat{r}_{i-1,j}$, $\hat{r}_{i,j-1}, \hat{r}_{i-1,j-1}$ by (6). At backward propagation, $\mu_{i,j}$ is calculated by (8). It gains the gradient from three directions proportional to how optimal the cumulative cost $\hat{r}$ of each direction is. Then, $\hat{\mu}_{i,j}$ is calculated from $\mu_{i+1,j}$, $\mu_{i,j+1}$, and $\mu_{i+1,j+1}$ proportional to how optimal each pairwise cost δ is in (10).

and the other blue and red terms are calculated in the same way. Like (9) which measures how minimal the $\hat{r}_{i,j}$ is among three directions, (11) measures how minimal the $\delta_{i,j}$ is among three directions. Hence, (8) aggregates global optimal path information and (10) aggregates local optimal path information due to the $\hat{r}_{i,j}$ at the former one and the $\delta_{i,j}$ at the latter one. Unlike S2DTW, the Soft-DTW only requires to calculate M matrix with r instead of $\hat{r}$ by (8) and then does not consider the local optimality. Figure 3 depicts the forward and backward propagation of S2DTW.

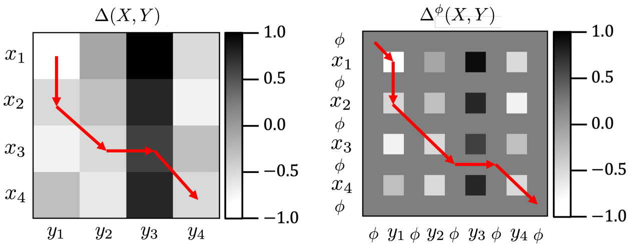
4.2. Weak Alignment

We further modify the Soft-DTW by allowing its path not to forcibly align irrelevant pairs as (Figure 1c and 1d). Besides, our S2DTW can start from (or end at) an arbitrary point. Adopting the trick in DWSA [42] for one-to-one matching with skipping, we achieve weak alignment by inserting dummy elements ϕ in the intervals (and both ends) of clip and caption sequences, (*e.g.*, $X = [x_1, x_2, \ldots, x_n]$ becomes $X^{\phi} = [\phi, x_1, \phi, x_2, \phi, \ldots, \phi, x_n, \phi]$).

In S2DTW, the pairwise distance matrix with dummy elements is $\Delta^{\phi}(X,Y) \in \mathbb{R}^{(2n+1)\times(2m+1)}$ and has dummy distance δ^{ϕ} at the pair which includes ϕ. δ^{ϕ} is a hyperparameter that can be interpreted as a threshold. By calculating the DTW with dummy elements, it leads the DTW path to pass only the pair whose distance is smaller than δ^{ϕ}. Unlike the standard DTW or Soft-DTW which forcibly align at least one pair per one timestamp, our proposed S2DTW weakly aligns the irrelevant clip-caption pairs and even enable many-to-many matchings which cannot be handled by DWSA. Figure 4a and 4b show the pairwise distance before/after adding dummy elements. This weak alignment framework is followed by the local neighborhood smoothing. As a result, the final pairwise distance is $\hat{\Delta}^{\phi}(X,Y)$ which is used to calculate the DTW.

4.3. Temporal Data Augmentation

As discussed in Section 3.2, videos often have non-sequential alignments, but the standard DTW cannot resolve them since it allows only three moves $\{\downarrow, \searrow, \rightarrow\}$. To address this problem, we propose a simple data augmentation that temporally shuffles clips and captions. Let π denote a permutation and then a clip permuted by π is $X_{\pi} = [x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}]$. To avoid temporally or semantically too extreme augmentations, we consider a subset of possible permutations. We first leave out the cases when a clip is temporally shifted beyond a time window. For example of $\forall j \in [1,n]$, the j-th clip cannot be out of the window of size w, *i.e.*, the range of possible indices after a permutation of j-th clip is $[\max(1, j-w), \min(n, j+w)]$. The set of permutations that satisfies this temporal constraint is



(a) Original Pair-wise Distance (b) Pair-wise Distance with Dummy

Figure 4. **Illustration of Weak Alignment.** (a): The original pairwise distance without dummy elements has to pass the pair whose clip and caption are irrelevant each other, *e.g.*, the caption y_3 is irrelevant with any other clips. (b): On the other hand, the pair whose distance is bigger than dummy distance can be skipped.

denoted as $\mathcal{T}(n, w)$. Given the temporal constraint, we propose the target distribution as follows:

$$p_{\pi}^{clip} = \begin{cases} \sigma\left(-\frac{\|\Delta(X,X)-\Delta(X_\pi,X_\pi)\|_2^2}{\tau}\right) & \text{if } \pi \in \mathcal{T}(w,n) \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where σ is softmax function computed over all permutations in $\mathcal{T}(w, n)$ and τ is a temperature parameter. $\Delta(X, X)$ and $\Delta(X_\pi, X_\pi)$ are self-similarity matrices before/after permutation. The proposed target distribution more likely generates a permutation that less changes the self-similarity structure. In other words, the proposed augmentation less likely generates semantically too strong augmentations that hinder representation learning. Then, the temporally augmented $\tilde{X} \sim P^{clip}(X; \Pi)$ which is a shuffled sequence of clips is sampled from the distribution P^{clip} defined in (12). The captions $\tilde{Y}$ is augmented in the same way and finally we calculate the pairwise distance matrix $\Delta(\tilde{X}, \tilde{Y})$ as the input for alignment (*e.g.*, DTW). For simplicity of implementation, each modality is shuffled independently.

Our temporal augmentation encourages learning invariant features under permutation and allow minimizing the distance between clips and captions that cannot be aligned by sequential alignment algorithms such as the standard DTW. This is helpful to learn representation when the clips and captions are non-sequentially aligned as in Figure 1b.

4.4. Contrastive Learning with S2DTW

With S2DTW, we perform representation learning in a self-supervised manner. S2DTW can be used for a distance measure between clips and captions. Minimizing the distance between two samples without negative pairs causes feature collapsing. Hence, to address this problem, we adopt a well-known contrastive loss, InfoNCE loss [38]. Our final loss is defined as:

$$\mathcal{L} = -\log \sum_i \left(\frac{e^{-\text{S2DTW}\left(X^{(i)},Y^{(i)}\right)}}{e^{-\text{S2DTW}\left(X^{(i)},Y^{(i)}\right)} + \sum_{j\in\mathcal{N}_i} e^{-\text{S2DTW}\left(X^{(i)},Y^{(j)}\right)}} \right) \tag{13}$$

, where $X^{(i)}$ and $Y^{(i)}$ are clips and captions of the i-th video and $\mathcal{N}_i$ is a set of negative samples of the i-th video in mini-batch. This formulation also implicitly mines the hard negatives. In a clip-caption level, due to the nature of the DTW, a clip-caption pair which has closer distance in negative samples will get stronger negative signal to push away than others in negative samples. Therefore, unlike in baseline [25], no additional hard negative mining strategy (*e.g.*, [23]) was taken for proposed method. Further discussions with qualitative results are in the appendix.

5. Experiments

In this section, we evaluate the performance on various downstream tasks by applying our pretrained feature embeddings (Section 5.1). We also describe ablation studies about the effect of each algorithm which is addressed in Section 4 and finally analyze qualitative results of each algorithm in terms of the DTW path (Section 5.2). All downstream tasks and ablation studies except for the action recognition task are conducted in the zero-shot learning setting to evaluate only the quality of learned representations. For the action recognition task, we adopt widely used linear evaluation protocol, which trains a linear classifier on top of the frozen representation. The experimental setup and further ablation studies are in the appendix.

5.1. Downstream Tasks

5.1.1 Action Recognition

We firstly evaluate learned video representation without using text representation on the action recognition task whose goal is to distinguish video-level actions. In Table 1, we compare the proposed method with other self-supervised methods. According to the linear evaluation protocol, our VT-TWINS outperforms all self-supervised learning methods including the baselines that performed fine-tuning denoted by (Frozen x) such as CBT [44] and 3DRotNet [24]. This result shows that our method improves the generality of video representations. Especially for HMDB, VT-TWINS obtains about 4% improvement over the MIL-NCE with the same backbone model (S3D).

5.1.2 Video and Text Retrieval

We evaluate the effectiveness of the joint representation of video and text by applying text-to-video and video-to-text retrieval tasks, which aim to find a corresponding clip (caption) given a query caption (clip).

Method	Dataset	MM	Model	Frozen	HMDB	UCF
OPN [1]	UCF	✖	VGG	✖	23.8	59.6
Shuffle & Learn [37]*	K600	✖	S3D	✖	35.8	68.7
Wang *et al.* [48]	K400	Flow	C3D	✖	33.4	61.2
CMC [47]	UCF	Flow	CaffeNet	✖	26.7	59.1
Geometry [16]	UCF	Flow	CaffeNet	✖	26.7	59.1
Fernanado *et al.* [15]	UCF	✖	AlexNet	✖	32.5	60.3
ClipOrder [54]	UCF	✖	R(2+1)D	✖	30.9	72.4
3DRotNet [24]*	K600	✖	S3D	✖	40.0	75.3
DPC [21]	K400	✖	3D-R34	✖	35.7	75.7
3D ST-puzzle [26]	K400	✖	3D-R18	✖	33.7	65.8
CBT [44]	K600	✖	S3D	✔	29.5	54.0
CBT [44]	K600	✖	S3D	✖	44.6	79.5
AVTS [28]	K600	Audio	I3D	✖	53.0	83.7
MIL-NCE [35]	HTM	Text	I3D	✔	54.8	83.4
MIL-NCE [35]	HTM	Text	S3D	✔	53.1	82.7
VT-TWINS	HTM	Text	S3D	✔	**57.9**	**85**
S3D (supervised learning) [53]			S3D	✖	75.9	96.8

Table 1. **Action Recognition.** Shuffle & Learn* and 3DRotNet* are reimplemented by [44] with S3D.

Method	Labeled Dataset	R@1	R@5	R@10	MedR
Random Init	None	0.03	0.15	0.3	1675
HGLMM FC CCA [27]	IM, K400, YC2	4.6	14.3	21.6	75
Miech *et al.* [36]	IM, K400	6.1	17.3	24.8	46
Miech *et al.* [36]	IM, K400, YC2	8.2	24.5	35.3	24
COOT [17]	YC2	5.9	16.7	24.8	49.7
ActBERT [58]	YC2	9.6	26.7	38.0	19
MIL-NCE [35]	None	8.8	24.3	34.6	23
VT-TWINS	None	**9.7**	**27**	**38.8**	**19**

Table 2. **Text-to-Video Retrieval on YouCook2.**

Text-to-video retrieval. Table 2 and 3 show the performance of text-to-video retrieval on YouCook2 and MSR-VTT dataset. For fair comparison with MIL-NCE, we trained our model on HowTo100M dataset and evaluate on the test set *without* any additional supervision. Table 2 shows that our VT-TWINS outperforms MIL-NCE and even other methods (*e.g.*, COOT and ActBERT) that are fine-tuned on YouCook2 (denoted as YC2). Similarly, on MSR-VTT dataset Table 3 shows that the proposed method outperforms several multi-modal self-supervised methods trained on the HowTo100M (MIL-NCE, Amrani *et al.*, SSB). In addition, our method is better or on par with ActBert that is fine-tuned on the target dataset MSR-VTT.

Video-to-text retrieval. We also compare the performance of video-to-text retrievals with MIL-NCE. Table 4 shows that our VT-TWINS outperforms MIL-NCE on both YouCook2 and MSR-VTT. Note that MIL-NCE blindly and equally treats all the captions in a time window around a query clip as positives. We believe that this assumption often does not hold and learning with the inaccurate clip-caption pairs may hinder learning representations to precisely associate clips and captions.

Method	Labeled Dataset	R@1	R@5	R@10	MedR
Random Init	None	0.01	0.05	0.1	500
Miech *et al.* [36]	IM, K400	7.5	21.2	29.6	38
Amrani *et al.* [2]	None	8.0	21.3	29.3	33
SSB [39]	None	8.7	23.0	31.1	31.0
ActBERT [58]	MSRVTT	8.6	**23.4**	**33.1**	36
MIL-NCE [35]	None	8.2	21.5	29.5	40
VT-TWINS	None	**9.4**	**23.4**	31.6	**32**

Table 3. **Text-to-Video Retrieval on MSRVTT.**

Method	YouCook2				MSRVTT			
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR
Random Init	0.03	0.13	0.26	1717.5	0.1	0.49	0.98	499.5
MIL-NCE* [35]	9.35	26.22	37.36	22	8.9	20.65	27.2	46
VT-TWINS	**9.7**	**28**	**40.3**	**16**	**9.1**	**22.9**	**29.1**	**43**

Table 4. **Video-to-Text retrieval.** * is our reproduction of official code of the MIL-NCE.

Method	Labeled Dataset	CTR
Alayrac *et al.* [1]	IM, K400	13.3
CrossTask [59]	IM, K400	22.4
CrossTask [59]	IM, K400, CT	31.6
Miech *et al.* [36]	IM, K400	33.6
DWSA [42]	CT	35.5
ActBERT [58]	CT	37.1
MIL-NCE [35]	None	35.5
VT-TWINS	None	**40.7**

Table 5. **Action Step Localization on CrossTask.**

5.1.3 Action Step Localization

We also evaluate the representations learned by our method in the action step localization task on the CrossTask dataset. We adopted the zero-shot evaluation suggested in [36]. Table 5 shows that VT-TWINS significantly outperforms baselines achieving an CrossTask average recall (CTR) of 40.7%. This surpasses MIL-NCE (35.5%) and even the models that are trained on the CrossTask dataset such as DWSA (35.5%) and ActBERT (37.1%).

5.2. Ablation Study and Qualitative Analysis

5.2.1 Temporal Data Augmentation

As explained in Section 4.3, the proposed augmentation less likely generates a permutation which is significantly different from the original sequence. To evaluate the effectiveness of our temporal data augmentation, we compare it with two other strategies: One is sampling from the uniform distribution and the other is sampling from a inverse distribution of our one, *i.e.*, assigning a higher probabil-

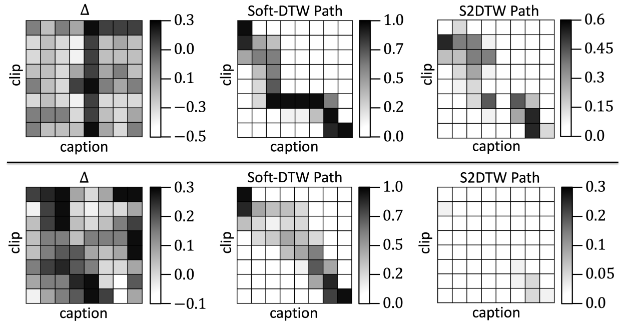


Figure 5. **Results of Weak Alignment.** Δ is a pairwise distance matrix. The Soft-DTW path and the S2DTW path matrices are the gradient matrices M and $\hat{M}$ defined in Section 4.1. Each row shows the partially and entirely irrelevant pairs, respectively.

ity to the semantically similar permutation with the original sequence. (2), (3), and (4) in Table 6 demonstrate that temporal shuffles while maintaining semantic information helps to learn feature representation on weakly correlated data with non-sequential alignments. Especially, the gap is substantial in the task that uses joint embedding representations (YouCook2, MSR-VTT, and CrossTask in Table 6) because strong augmentation harms semantic information a lot, it is difficult to learn the representations aligned between clip-captions.

5.2.2 Weak Alignment

The top row of Figure 5 shows the case of partially irrelevant pairs; the pairwise distance matrix Δ on top shows that the fifth caption has a consistently large distance from the other clips[1]. In this case, the Soft-DTW is enforced to align one or more pairs per each timestamp. On the other hand, S2DTW shows the results that the unrelated pairs are weakly aligned because S2DTW skips them appropriately. Moreover, the Soft-DTW has another problem that it is forced to align the start point (1,1) and the end point (n,m). Unlike the Soft-DTW, we observe that S2DTW can ignore the start point and the end point.

The Soft-DTW also finds a temporal alignment path even in the entirely uncorrelated data like the case of Figure 1d. The bottom row of Figure 5 illustrates that most elements of the pairwise distance are greater than zero (the leftmost matrix), *i.e.*, the clips and captions are almost entirely irrelevant. The path is clearly drawn in the Soft-DTW while most elements are not learned in S2DTW by aligning weakly. (2) and (5) of Table 6 show that weak alignment of S2DTW improves the performance on the weakly correlated data by ignoring irrelevant pairs.

	TA	WA	LS	HMDB	UCF	YC2	MV	CT
(1)	-	-	-	38.9	68.6	8.7	12.7	22.9
(2)	**A**	-	-	39.4	69.3	9.6	13.6	23.5
(3)	**B**	-	-	36	68.5	5	10.5	17.4
(4)	**C**	-	-	36.9	68	4.9	11.5	16.8
(5)	**A**	✔	-	39.1	70.6	10.6	14.7	26.9
(6)	**A**	✔	✔	42	72.1	12.5	17.4	28.2

Table 6. **Ablation Studies.** We report accuracy on the **HMDB** and **UCF**, R@10 on the YouCook2 (**YC2**) and MSR-VTT (**MV**), and CTR on the CrossTask (**CT**) to evaluate the contribution of the followings: temporal data augmentation (**TA**), weak alignment (**WA**), and local neighborhood smoothing (**LS**). Each element is applied to the standard DTW with a contrastive learning scheme. For **TA**, we evaluate the following strategies: **A**: suppressing semantically strong permutation (ours), **B**: random permutation, and **C**: encouraging semantically strong permutation (inverse of ours).

5.2.3 Local Neighborhood Smoothing

We also evaluate the effectiveness of local neighborhood smoothing. As mentioned in Section 4.1, local neighborhood smoothing can reflect local optimal path as well as global optimal path. (5) and (6) in Table 6 show that local neighborhood smoothing complements the DTW and improves the performance.

6. Conclusion

We have presented a novel multi-modal self-supervised learning framework for learning joint embeddings of video and text from uncurated narrated videos. To address the challenges of weakly correlated video and caption pairs, our framework VT-TWINS first aligns the clips and captions by the proposed weak alignment algorithm and learns representations via contrastive learning. Our experiments on a wide range of three tasks over five benchmark datasets demonstrate that the proposed method significantly improves the generality of joint embeddings and outperforms self-supervised methods as well as fine-tuned models on target tasks. The proposed framework is a generic framework that is applicable in representation learning with multi-modal time-series data. Future directions, limitations, and negative societal impacts are discussed in the appendix.

Acknowledgments. This work was partly supported by Efficient Meta-Learning Based Training Method and Multipurpose Multi-Modal Artificial Neural Network for Drone AI (No.2021-0-02312), and ICT Creative Consilience program (IITP-2022-2020-0-01819) supervised by the IITP; the National Supercomputing Center with supercomputing resources including technical support (KSC-2021-CRE-0299) and Kakao Brain corporation.

[1] Also refer to Figure 1c as an illustrative example.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 2, 7

[2] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *AAAI*, 2021. 7

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2

[5] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, 1994. 2

[6] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *CVPR*, 2019. 2

[7] Xiaobin Chang, Frederick Tung, and Greg Mori. Learning discriminative prototypes with dynamic time warping. In *CVPR*, 2021. 2

[8] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *ICCV*, 2021. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 2

[11] Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev, and Cordelia Schmid. A flexible model for training action localization with varying levels of supervision. In *NeurIPS*, 2018. 1

[12] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *ICML*, 2017. 2, 3, 4

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1

[14] Nikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan D Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. In *NeurIPS*, 2021. 2

[15] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017. 1, 2, 7

[16] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, 2018. 7

[17] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. In *NeurIPS*, 2020. 2, 7

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 2

[19] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 1

[20] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *CVPR*, 2021. 2

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV*, 2019. 2, 7

[22] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *CVPR*, 2021. 2

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 6

[24] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations, 2018. 2, 6, 7

[25] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, 2020. 6

[26] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019. 2, 7

[27] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015. 1, 7

[28] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 2, 7

[29] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *ECCV*, 2016. 1

[30] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 2

[31] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 1

[33] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation, 2020. 2

[34] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1

[35] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 1, 2, 3, 7

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1, 2, 3, 7

[37] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016. 2, 7

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. 6

[39] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 7

[40] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 2

[41] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE TrASSP*, 1978. 1, 2

[42] Yuhan Shen, Lu Wang, and Ehsan Elhamifar. Learning to segment actions from visual and language instructions via differentiable weak sequence alignment. In *CVPR*, 2021. 5, 7

[43] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *CVPR*, 2017. 1

[44] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer, 2019. 2, 6, 7

[45] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[46] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 1

[47] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 7

[48] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019. 7

[49] Jinpeng Wang, Yiqi Lin, Andy J Ma, and Pong C Yuen. Self-supervised temporal discriminative learning for video representation learning, 2020. 2

[50] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE TPAMI*, 2018. 1

[51] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1

[52] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 2019. 1

[53] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 7

[54] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019. 1, 2, 7

[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1

[56] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 1

[57] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1

[58] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 2, 7

[59] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 7