TEMPCLR: TEMPORAL ALIGNMENT REPRESENTATION WITH CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video representation learning has been successful in multi-modal pretraining where each sentence in a paragraph of description is trained to be close to the paired video clips in the common feature space. For long videos, in particular, by optimizing the sentence-clip pairs, the full video and its text description are aligned implicitly. However, such strict optimization may ignore the temporal context over a long time span, which inevitably limit the generalization ability. In this paper, we propose a contrastive learning framework, TempCLR, and compare the full video and paragraph directly. As the video (paragraph) can be formulated as a sequence of clips (sentences), we first calculate pair-wise sentence-clip matching cost. Then, under the constraint of temporal ordering, we use dynamic time warping to calculate the minimum alignment cost, *i.e.*, distance, between video and the corresponding paragraph. To explore the temporal dynamics, we break the temporal consistency by shuffling the video clips or sentences separately and then find the distance to be maximized. In this way, we learn to extract features for clips and sentences which are aware of the global temporal information and are then friendly for sequence alignment. In addition to video-paragraph alignment, our approach can also be generalized on the matching between different short video instances. We conduct experimental study on action step localization, video retrieval, and few-shot action recognition, and achieved consistent performance gain over all three tasks. Detailed ablation studies are provided to justify the selection of each component.

1 INTRODUCTION

Representation learning on videos has achieved remarkable success (Goroshin et al., 2015; Feichtenhofer et al., 2021) and has been extended on video-text data (Miech et al., 2019; Radford et al., 2021) to learn a common feature space for zero-shot transfer. Given a paragraph of text description for the full video, in addition to recognizing actions in a short video, the understanding of long videos is also increasingly important such as [placeholder: example of long video understanding].

A long video is usually formulated as a sequence of short video clips. Given a paragraph, every sentence is used to describe (*i.e.*, paired with) the consecutive video clips in a temporal segment. Then, by training each sentence to be close to the corresponding clips *locally* Miech et al. (2020), the video and its paired paragraph can be aligned implicitly. Similarly, by fusing the embeddings over a short time span for text and video clips in advance, Xu et al. (2021) learns to align the fused embeddings but still does not directly align the video and paragraph. Instead, as a paragraph is essentially a sequence of sentences, by modeling the correlation between the two sequences (sequence-level), the whole long video and the paragraph can also be compared *globally* (Fig. 1(a)).

In addition, obtaining good matching between individual clips and sentences (unit-level) ignores the context of temporal dynamics, and will limit the generalization ability (Goyal et al., 2017). For all clips within a video segment, as the action/event progress at each clip varies, strictly matching the sentence with all of the paired clips during pretraining, serving as a *hard-label*, may not always result in an optimal solution. Also, as the visual contents are mostly dominated by background, the clips in different temporal segments can be naturally similar. For a challenging case where two clips are visually similar (*i.e.*, confusing) but are supposed to be paired with different sentences (Fig. 1(b)), the context information by temporal order can avoid the potential mismatching in unit-

Figure 1: Previous methods assume XXX. However, directly modelling the distance between them will be also important.

level comparison. Similarly, for the segment features which are obtained by aggregating the related clip features, the segment feature may still be dominated by spatial information (Buch et al., 2022) and the temporal understanding may still be less explored during training.

In this work, we propose a contrastive learning framework to compare the full video and paragraph directly and explore temporal dynamics TempCLR. We directly calculate the distance between full video and paragraph during comparison, and the distance is the minimum cumulative matching cost over the sentences and clips under the constraint of temporal dynamics. Specifically, for the paragraph (anchor) and its paired video (positive), we calculate the pairwise distance between sentences and clips and then use dynamic time warping (DTW) (Müller, 2007) to find the distance. Then, [to integrate the temporal context into embedding], we emphasize the relation between units within the same sequence and consider the cases where the consistency between video and paragraph is not met. Without loss of generality, for each paragraph, we shuffle the order of clips in the paired video to synthesize negative samples and then use DTW to find the minimum distance to be maximized. Finally, we apply contrastive learning (Chen et al., 2020a) to maximally align video and paragraph in a positive pair. In this way, we can learn embeddings for clips and sentences which are aware of temporal context in a long time span and are friendly for sequence matching.

In addition to compare sequences for vision-language pretraining, our TempCLR can also be generalized on one-shot action recognition where different videos are compared (*i.e.*, video-only). During the evaluation, each video is classified through the nearest neighborhood search according to the sequence distance. As the duration of each action is short, we formulate each video as sequence of frames. As the frame matching between videos are not annotated, we also replace DTW with OTAM due to a strong assumption in DTW Cao et al. (2020) to calculate the distance instead. In summary, the contributions are:

- We proposed a contrastive learning framework TempCLR to directly compare sequences and the learned representation for clips and sentences perceive temporal dynamics.
- Given an anchor, we designed a negative sampling strategy by shuffling the positive sequence and used DTW to directly calculate sequence-wise distance. Notably, our method can be generalized to learn representation for both video-paragraph data and video-only data.
- We conducted extensive experiments on three tasks (*i.e.*, action step localization, video retrieval, and few-shot action recognition) to demonstrate the effect of our training strategy. At each task, we observed consistent performance gain and achieved state-of-the-art performance. Comprehensive ablation studies are provided to justify the design of each component.

2 Related Work

Contrastive learning has achieved remarkable success on images (Chen et al., 2020b; He et al., 2020) and videos (Feichtenhofer et al., 2021; Recasens et al., 2021). The main idea is to group different views of the same image/video instance, and push away negative samples from different instances by minimizing the InfoNCE loss (Oord et al., 2018) and it . It uses cosine similarity during feature comparison and can cluster features even though no semantic class labels are not provided (Khosla et al., 2020), which is also theoretically analyzed in Wang & Isola (2020). (Caron et al., 2020) then improves the feature clustering on imbalanced datasets. Besides, by using different backbones to synthesis features for comparison, the image representation can also be improved (Grill

et al., 2020; Ma et al., 2021). In addition to the spatial augmentation, the temporal information in videos can also be exploited. Feichtenhofer et al. (2021) assumes the visual content of all clips from the same video are consistent when the length is short. Then, Recasens et al. (2021) generates views by using different temporal length. Meanwhile, Jenni et al. (2020) comprehensively studies the effect of different temporal augmentation methods for contrastive learning.

Multi-modal pre-training for zero-shot transfer has been has been studied to connect vision and language. CLIP (Radford et al., 2021) applied contrastive learning on image-caption pairs and learns a common multi-modal feature space which are transferable for new tasks. Then, Yang et al. (2022) and Li et al. (2022) further modified the negative sampling strategy such that the embeddings can be more discriminative and can be used for object detection separately. Similarly, such pre-training has also been extended to video understanding Miech et al. (2020), and the pioneering works include Ging et al. (2020); Gabeur et al. (2020); Alayrac et al. (2020). Multi-task pretraining has been studied to improve the performance (Li et al., 2020; Luo et al., 2020). Then, to mitigate the impact of noise in long videos labeling, VideoCLIP (Xu et al., 2021) proposed a sampling strategy and improved the performance. Besides video and text, the audio can also be included to benefit the zero-shot tasks on long-videos (Chen et al., 2021; Shvetsova et al., 2022). As an alternative, instead of learning a feature space, the flexibility of the transformer structure Vaswani et al. (2017) can also be utilized where the cross-attention mechanism can fuse the multi-modal information at each layer automatically (Sun et al., 2019; Su et al., 2019; Zhu & Yang, 2020; Chen et al., 2020c).

Sequence alignment has been well-studied in conventional machine learning. For each unit in the sequence/time-series, under the constraint of temporal ordering, the indexes of matched units from the aligned sequence should be monotonically increasing. Then, an optimal matching between aligned sequences can be found where the averaged distance over matched units is minimized. Dynamic time wrapping (DTW) Müller (2007) is first proposed to find the optimal matching though dynamic programming where canonical time warping Zhou & Torre (2009) is then used to align sequences with different feature dimensionality and is applied in deep learning methods (Sargin et al., 2007). Meanwhile, a pyramid deep architecture (Wang et al., 2020) or attention-based mechanism (Bishay et al., 2019; Zhang et al., 2020) can be designed to integrate multi-scale temporal information into a single feature vector for comparison. Besides, under the regularization of the sequence alignment, several pre-training strategies are designed for visual-audio and visual-rhythms synchronization (Cheng et al., 2020; Yu et al., 2022) as well as video-text alignment Xu et al. (2021). However, all of the training objectives are still focus on the matching between units and assume the sequences can be aligned implicitly. Though Huang et al. (2021a) has considered on the order between two related actions for action localization, it does not consider the alignment with the full video and thus does not explore the complete temporal context.

3 Approach

We first provide notation and task formulation in Sec. 3.1. Then, we detail the paradigm of our method and explain how to adapt it for different tasks in Sec. 3.2 and 3.3 respectively.

3.1 Pre-Training Task Formulation

Given an anchor instance $\mathbf{S}_a$ (*i.e.*, a paragraph or a video), we aim to learn a network that can minimize its distance with a positive instance $\mathbf{S}_p$ (*i.e.*, a video paired with the paragraph or a video of the same semantic class). Since the time span for each paragraph (video) can be long, it is typically formulated as a sequence of sentences (short video clips). Then, a network is trained to extract a feature embedding for each sentence (short video clip), resulting in a sequence of embeddings, *i.e.*, $\mathbf{S}_a = \{\mathbf{s}_a^i\}_{i=1}^{N_a}$ and $\mathbf{S}_p = \{\mathbf{s}_p^j\}_{j=1}^{N_p}$, where the intrinsic temporal order within $\mathbf{S}_a$ and $\mathbf{S}_p$ are consistent with each other ($\mathbf{s}_a^i, \mathbf{s}_p^j \in \mathcal{R}^d$ are the sequence units where d is the dimension of the common feature space, N_a and N_p are the sequence lengths but N_a is not necessarily equal to N_p). In this way, to directly calculate the distance $d_{\{\mathbf{S}_a,\mathbf{S}_p\}}$, we use the minimum alignment cost over units between two sequences $\mathbf{S}_a$ and $\mathbf{S}_p$, where $d_{\{\mathbf{S}_a,\mathbf{S}_p\}}$ is supposed to be small when $\mathbf{S}_a$ and $\mathbf{S}_p$ are aligned.

Dynamic Time Wrapping (DTW) is introduced by Müller (2007) and aims to find an optimal alignment over units between sequences where the sum of costs over matched units is minimum. We set a matching matrix $M \in \mathcal{R}^{N_a \times N_p}$ where $M(i,j) = 1$ indicates $\mathbf{s}_a^i$ and $\mathbf{s}_p^j$ are matched. Then,

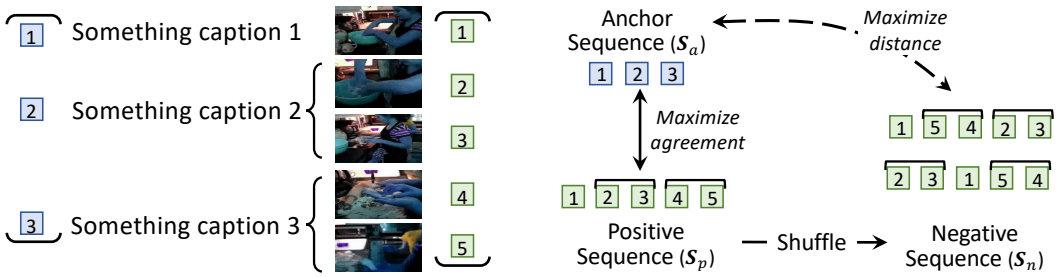

Figure 2: Illustration of our approach. Without loss of generality, we first extract the embedding for each caption (blue) and each clip (green) to build the anchor $\mathbf{S}_a$ and positive $\mathbf{S}_p$ sequences [update caption]. Then, we generate negative samples by shuffling the embeddings in $\mathbf{S}_p$ but still keep the order. Then, for the anchor, the training objective is to maximize its agreement with the positive one and maximize its distance with each of the negative sequences.

every $\mathbf{s}_p^j$ in $\mathbf{S}_p$ is supposed to be matched with at least one unit in $\mathbf{S}_a$, *i.e.*, $\sum_{1\leq i\leq N_a} M(i,j) \geq 1$ for $\forall j \in \{1...N_p\}$, and vice versa. Under the constraint of sequence order, the index of matched unit should be monotonously increasing, *i.e.*, if $\mathbf{s}_p^n$ matches $\mathbf{s}_a^m$ where $1 \leq m \leq N_a$ and $1 \leq n < N_p$, $\mathbf{s}_p^{n+1}$ cannot be matched with any unit in $\{\mathbf{s}_a^i\}_{i=1}^{m-1}$, *i.e.*, $\sum_{1\leq i<m} M(i,j) = 0$.

Firstly, DTW calculates the pair-wise matching costs between units $D \in \mathcal{R}^{N_a \times N_p}$ where $D(i,j)$ is the cost, *e.g.*, cosine distance (Singhal et al., 2001), between $\mathbf{s}_a^i$ and $\mathbf{s}_p^j$. Then, DTW employs a dynamic programming strategy and sets a matrix $C \in \mathcal{R}^{N_a \times N_p}$ to record the minimum cumulative cost until $\mathbf{s}_a^i$ and $\mathbf{s}_p^j$ (Wang et al., 2020; Dixit et al., 1990), *i.e.*,

$$C(i,j) = D(i,j) + \min\{C(i-1,j-1),\ C(i-1,j),\ C(i,j-1)\}. \tag{1}$$

Then, the sequence distance is $d_{\{\mathbf{S}_a,\mathbf{S}_p\}} = C(N_a, N_p)$ and M can be recorded at the same time.

3.2 Temporal Alignment Representation with Contrastive Learning

The temporal dynamics naturally exhibited in each sequence (*i.e.*, paragraph or video) are representative. As $\mathbf{S}_a$ and $\mathbf{S}_p$ are of similar semantic meaning and their temporal orders are consistent, when a global optimal alignment between $\mathbf{S}_a$ and $\mathbf{S}_p$ is found, the units $(\mathbf{s}_a^i, \mathbf{s}_p^j)$ in each matched pair are also assumed to be semantically close to each other. As such, when the features in $\mathbf{S}_p$ are hard to be distinguished due to redundant information such as visual background, the network may utilize the global temporal order to find a proper alignment between $\mathbf{S}_a$ and $\mathbf{S}_p$.

In contrast, when the temporal consistency w.r.t. $\mathbf{S}_a$ is not preserved, *i.e.*, a negative sequence $\mathbf{S}_n$ whose semantic meaning is different or temporal order is not consistent with $\mathbf{S}_a$, the distance $d_{\{\mathbf{S}_a,\mathbf{S}_n\}}$ between $\mathbf{S}_a$ and $\mathbf{S}_n$ should be high. After all, $\mathbf{S}_a$ cannot be aligned with $\mathbf{S}_n$, and the units in every matched pair derived by DTW are not guaranteed to be of high correlation.

Specifically, we consider the importance of temporal order and aim to obtain feature embeddings that preserve the order information and thus can benefit the alignment between two sequences. Thus, as shown in Fig. 2, we propose TempCLR to learn representations that can facilitate the temporal alignment with $\mathbf{S}_a$. We first generate negative sequences by randomly shuffling the units within $\mathbf{S}_p$ and then follow the contrastive learning (Chen et al., 2020a) framework.

Contrastive learning (CT) is a self-supervision approach by learning to group the samples with high correlation. CT treats different views of the same instance as correlated (positive) and builds negative pairs by sampling views from different instances. As each instance is an image in (Chen et al., 2020a), each view can be generated by performing augmentation on the low-level pixels, *i.e.*, adding effects such as flipping and color distortion. In practice, given a set $\mathcal{B}$ with N_B instances, for each instance $I \in \mathcal{B}$, another view I' is generated and is then used to build a positive pair with I where the other instances in $\mathcal{B}$ are used to build negative pairs with I. Then, the training objective is to minimize the InfoNCE loss Oord et al. (2018),

$$\mathcal{L}_{CT}(I, I', \mathcal{B}_n) = -\log \frac{\exp(\mathbf{z}_I \cdot \mathbf{z}_{I'}/\tau)}{\exp(\mathbf{z}_I \cdot \mathbf{z}_{I'}/\tau) + \sum_{X\in\mathcal{B}_n} \exp(\mathbf{z}_I \cdot \mathbf{z}_X/\tau)}. \tag{2}$$

where $\mathcal{B}_n = \mathcal{B} \setminus \{I\}$, $\mathbf{z}_X \in \mathcal{R}^d$ is the feature for instance X after l_2-normalization and τ is a hyperparameter to rescale the affinity scores. In this way, CT is performing pair-wise comparison where the disagreement between the two features in a positive pair is induced by the variation of augmentation. Then, for our approach, we can derive the training objective, *i.e.*,

$$\mathcal{L}_{seq}(\mathbf{S}_a, \mathbf{S}_p, \mathcal{S}_n) = -\log \frac{\exp(d_{\{\mathbf{S}_a, \mathbf{S}_p\}}/\tau)}{exp(d_{\{\mathbf{S}_a, \mathbf{S}_p\}}/\tau) + \sum_{\mathbf{S}_n \in \mathcal{S}_n} \exp(d_{\{\mathbf{S}_a, \mathbf{S}_n\}}/\tau)} \tag{3}$$

where $\mathcal{S}_n = \{\mathbf{S}_n^{(i)}\}_{i=1}^{N_n}$ is the set of N_n negative sequences derived from $\mathbf{S}_p$. As a complementary component, the sequences which are from other instances unpaired or uncorrelated with $\mathbf{S}_a$ can also be included in $\mathcal{S}_n$. However, it introduces more computation workload but does not improve performance (analyzed in Sec. 5.1) effectively. As such, for each $\mathbf{S}_a$, we only use the shuffling strategy for negative sample generation. By minimizing $\mathcal{L}_{seq}$, $\mathbf{S}_a$ and $\mathbf{S}_p$ are trained to be close to each other while both of them are pushed away from all negative sequences in $\mathcal{S}_n$.

3.3 Adaptation for Pretraining Tasks

We briefly explain how to apply the paradigm to align the video and paragraph in one semantically-correlated pair (video-paragraph) or videos of the same class (video-only) during network training.

Video-Paragraph. For each long video, a paired paragraph consisting of short sentences is provided and every sentence, *i.e.*, caption, describes the visual content within a temporal segment. A text encoder $\mathcal{F}_T : \mathcal{R}^{N_l \times d_w} \rightarrow \mathcal{R}^d$ is applied to extract embedding for each sentence, where $\{N_l, d_w\}$ are the number of words and the dimension of word vectors respectively. Then, for each video, all frames are grouped into non-overlapping consecutive short video clips and each clip, serving as a visual token, has n_f frames. Thus, each sentence is mapped with multiple consecutive clips. We use a visual encoder $\mathcal{F}_V : \mathcal{R}^{h \times w \times 3 \times n_f} \rightarrow \mathcal{R}^d$ to extract token embedding for each clip where $\{h, w\}$ denote the height and width of each frame.

For each sentence in the paragraph $\mathcal{S}_a$, we can find the clips in the annotated segment. By concatenating the clip embeddings for all sentences, we build $\mathcal{S}_p$. Since the segments of different sentences may have overlap, during training, without loss of generality, we will sub-select the sentences to have $\mathbf{S}_a$ such that there is no repeating clips in $\mathbf{S}_p$, *i.e.*, for each $\mathbf{s}_a^i \in \mathbf{S}_a$ with period $[t_0^i, t_1^i]$ where t_0^i and t_1^i are the starting and ending indexes of the clips in the full video, we always have $t_0^{i+1} > t_1^i$. Finally, to generate negative samples, we first change the order of the segments and then shuffle the clips within each segment.

Video-only. Our approach can be generalized to match different video instances of the same class, facilitating one-shot action recognition. Since the duration of action is short, we extract an embedding of each frame instead to formulate sequences. According to the starting and ending frame index as well as the sampling rate, we can determine the frames to be sampled. Then, we generate the negative sequences by directly shuffling the frames. Besides DTW, Cao et al. (2020) proposed OTAM, a variant of DTW, to avoid the restrict boundary constraint DTW, *i.e.*, $M(1,1) = M(N_a, N_p) = 1$. The effect of DTW and OTAM in our approach can be found in Sec. 5.2

4 Experiment

4.1 Evaluation Task

We evaluate our TempCLR on three task types, including action step localization, video retrieval, and one-shot action recognition.

Action step localization assumes that each video is associated with a Task consisting of multiple steps. Each step is explained in a specific video segment and is in a form of sentence description. Then, during inference, given a Task and its corresponding step condidates, each video clip is supposed to be assigned to the corresponding step. The performance is measured by recall, *i.e.*, number of step assignments that fall into the correct ground truth interval, divided by the total number of steps in the video.

Video Retrieval aims to find the matching video from a pool of videos, given the sentence query description. As the conventional setting is to matching the caption with a video clip (caption-clip),

Chen et al. (2021) recently formulates a more realistic scenario, full-video retrieval. Given a set of caption queries, *i.e.*, paragraph, describing multiple parts of an entire long video, the task then aims to retrieve the full video according to the paragraph. For both caption-clip retrieval and full-video retrieval, we use recall as metrics under different ranking patience, *i.e.*, R@1, R@5, R@10.

Since both action step localization and full-video retrieval consider the full video during evaluation. We mainly use these two tasks to examine the importance of utilizing the temporal order/dynamics and show how our approach benefit them.

One-shot action recognition aims to recognize action with only one labeled data. In each one-shot Task, given N_C classes where each class has one support video as a reference, we classify a test video by comparing its distance with each support video through the nearest neighbor search. In a common setting (Zhu & Yang, 2018), we are first given a *base* dataset of classes $\mathcal{C}_{base}$ to pre-train a model and a *novel* dataset of classes $\mathcal{C}_{novel}$ for evaluation. As the two class sets are *disjoint*, *i.e.*, $\mathcal{C}_{base} \cup \mathcal{C}_{novel} = \emptyset$, and no finetuning on support videos in *novel* is applied, the recognition accuracy can be used to indicate the generalization ability of our approach on new classes.

4.2 Dataset and Implementation Details

Video-Paragraph Pre-Training We follow Xu et al. (2021) and use HowTo100M (HT100M) (Miech et al., 2019) for network training. HowTo100M contains XXX instructional videos and the videos existing in YouCookII (Zhou et al., 2018) have been removed for fair comparison. Due to the computation resource limitation, we directly build our model on top of the VideoCLIP, *i.e.*, initialize the network parameters with the weights in a fully-trained VideoCLIP checkpoint, and randomly select 30k videos (2.7%) of the full training set to update the pre-trained network by minimizng our $\mathcal{L}_{seq}$. Among the subset, on average, the duration of each video is around 6.5 minutes with about 110 clip-caption pairs. The total text transcriptions is about xxx MB, with 2.4 tokens per second on average.

(*Architecture*) VideoCLIP (Xu et al., 2021) consists of two Transformer architectures for video and paragraph separately. For videos, a backbone network is used to extract embedding for each clip, *i.e.*, video token, where the clip embeddings are then fed into the 6-layer video transformer. For each sentence in the paragraph, they first obtain embedding for each token [though embedding look-up used in BERT (Devlin et al., 2018)] and then send the tokens of each sentence into a 12-layer Transformer. VideoCLIP uses S3D feature (Xie et al., 2018) as the clip embeddings and the S3D network has been pre-trained on HowTo100M (Miech et al., 2019) in a self-supervised manner (Miech et al., 2020). To align the dimension of embeddings, VideoCLIP additionally set a MLP with the Transformer to map the clip emedding dimension from 512 to 768. VideoCLIP is optimized by minimizng the InfoNCE (Chen et al., 2020a) between captions and clips, and does not clearly model the temporal orders during trianing. As VideoCLIP has already been a strong baseline, we choose it for comparison to better demonstrate the importance of temporal modelling.

For implementation, we use the script provided in VideoCLIP to extract the token embeddings for each video. During pre-training, we update all parameters of whole network, *i.e.*, two Transformer and one MLP layer. We train the model on 2 NVIDIA TiTAN RTX GPUs (each with 24 GB memory) for 10 epoches within two hours. We use the default hyper-parameters in Adam (Kingma & Ba, 2014) optimizer with betas of (0.9, 0.98) with a small learning rate $1e^{-5}$. More details can be found in Appendix.

Video-Paragraph Downstream Evaluation.
Action step localization is evaluted on CrossTask (Zhukov et al., 2019), which consists 83 different Tasks over 4.7K videos. [Each frame of video is annotated with one or multiple steps as a distribution.] We use the official testing data split which contains 1690 annotated videos over 18 Tasks and test set of CrossTask have been removed from the HowTo100M training set to avoid label leakage during pre-training. Firstly, we directly apply the model pre-trained on HT100M on CrossTask test set for zero-shot evaluation. Meanwhile, we can finetune the pre-trained VideoCLIP on the CrossTask train set, which contain 65 Tasks over XXX videos, and use the finetuned model for evaluation (*Supervised*). To note, there is no Task overlap between the train set and the test set and the step condidates in test set has never been seen during finetuning. For localization, for each video token, we compare its similarity with all step candidates after softmax normalization to have the confidence score.

Video Retrieval is evaluated on Youcook2 (Zhou et al., 2018) which consists of 2000 cooking videos with a total duration of 176 hours. On average, Each video is about 5.26 minutes and contains [XX capations]. We use the pre-trained model for zero-shot evaluation and there are in total 3305 caption-clip pairs from 430 videos. In full-video retrieval, for each video, Chen et al. (2021) directly concatenate the clips which have paird captions and assumes the background has been removed from the whole video. To mimic a more realistic scenario, we do not utilize the temporal annotation of clips and use the paragraph to retrieve the full video containing background.

Video-Only experiment is conducted on Something-Something V2 (Sth-Sth) by Goyal et al. (2017) and uses the subset provided by (Cao et al., 2020). The subset contains 100 classes and each class has 100 samples. The whole subset is then split into 64 *base* classes for pretraining, 24 (12) *novel* classes for evaluation (validation). During evaluation, following the protocol defined in Zhu & Yang (2018), we sample 15 test videos for each class in one Task and calculate the accuracy over $15N_C$ test samples. Finally, we report the mean accuracy over 1000 Tasks.
(*Architecture*) For fair comparison, we first use a ResNet50 model pretrained on ImageNet Deng et al. (2009) to extract embedding for each frame. Then, the embeddings of one video are fed into a 6-layer Transformer and the output features are $\mathbf{S}_a$. To obtain $\mathbf{S}_p$, we set a linear layer to process each frame embedding in $\mathbf{S}_a$ and the linear layer is jointly learned. In this way, no action label is used during training and the model is trained in a self-supervised manner.

4.3 Experimental Comparison

Table 1: Performance on action step localization for zero-shot (Left) and Supervised (right).

Approach (Zero-shot)	Recall
HT100M(Miech et al., 2019)	33.6
MIL-NCE(Miech et al., 2020)	40.5 (36.4)
MCN(Chen et al., 2021)	35.1
DWSA(Shen et al., 2021)	35.3
UniVL(Luo et al., 2020)	42.0
VT-TWINS(Ko et al., 2022)	40.7
VideoCLIP(Xu et al., 2021)	33.9
TempCLR (Ours)	36.9

Approach (Supervised)	Recall
Alayrac(Alayrac et al., 2016)	13.3
Zhukov(Zhukov et al., 2019)	22.4
Supervised(Zhukov et al., 2019)	31.6
VideoCLIP(Xu et al., 2021)	47.3
TempCLR (Ours)	**51.7**
Approach (Few-shot)	**Recall**
TempCLR (Ours) w/ 10%	40.7
TempCLR (Ours) w/ 20%	42.8

Action Step Localization As shown in Table 1, by adding the loss $\mathcal{L}_{seq}$, our method achieves consistent (9%) performance gain compared with the VideoCLIP baseline in both *zero-shot* and *Supervised*. Though VideoCLIP has shown strong performance after finetuning, our method can still keep increasing the performance from 47.3 to 52.5. Since VideoCLIP introduces a limited gain over MLP-NCE and we only use 2.7% training data for pre-training, the performance gain is not as high as SOTA 42.0 (R_5) but we still improve it from 33.9 to 36.9. Meanwhile, the two values in R_2 are obtained by using different feature embedings. However, by finetuning on 10 20% of training data of CrossTask, our approach can effectively improve the performance.

Table 2: Performance comparison on video retrieval (clip-caption)

Approach	backbone	R@1	R@5	R@10
Random	-	0.0	0.2	0.3
MIL-NCE*(Miech et al., 2020)	R152+RX101	8.1	23.3	32.3
MCN(Chen et al., 2021)	R152+RX101	18.1	35.5	45.2
MMV(Alayrac et al., 2020)	TSM-50x2	11.7	33.4	45.4
ActBERT(Zhu & Yang, 2020)	R101+Res3D	9.6	26.7	38.0
MIL-NCE(Miech et al., 2020)	I3D-G	11.4	30.6	42.0
HT100M(Miech et al., 2019)	S3D-G	6.1	17.3	24.8
MIL-NCE(Miech et al., 2020)	S3D-G	15.1	38.0	51.2
MMFT(Shvetsova et al., 2022)	S3D-G	**24.6**	48.3	60.4
VideoCLIP(Xu et al., 2021)	S3D-G	22.7	50.4	63.1
TempCLR(Ours)	S3D-G	23.3	**51.0**	**64.5**

Video Reterival. As shown in Table 3, we first follow MCN to measure the full-video retrieval performance while the background is removed. Since Chen et al. (2021) does not consider the temporal ordering, given a paragraph, MCN will use the each caption to retrieve each clip first and the video whose clips are mostly matched with the captions will be retrieved (Cap. Avg). Surprisingly, VideoCLIP has already achieves superior performance (R_4). Since the our approach is to model the temporal ordering within each sequence, TempCLR achieves similar performance (R_6). Then, we use DTW as measurement to compare the paragraph and full video directly. As VideoCLIP does not model temporal order, the performance is imilar to MCN ($R_{3,5}$). However, our approach can clearly improve the performance, which further demonstrate the importance of ordering in tmeporal modeling. Then, as shown in Table 6, we evaluate the retrieval of full video with background. By using different distance metric, our TempCLR can outperform the VideoCLIP baseline consistently. Finally, though our approach is not specifically designed to facilitate the matching on unit-level, as shown in Table 2, TempCLR can still benefit the matching between captions with clips. Since we are using the Transformer to process the clip embeddings in one video, we think the reason of the gain is that our $\mathcal{L}_{seq}$ loss can include global temporal information in the embedding, which also facilitate the matching. More explanation of implementation detail can be found in Appendix.

Table 3: Comparison on full-video retrieval.

Approach	Measure	R@1	R@5	R@10
MIL-NCE*	Cap. Avg.	43.1	68.6	79.1
HT100M*	Cap. Avg.	46.6	74.3	83.7
MCN(Chen et al., 2021)	Cap. Avg.	53.4	75.0	81.4
VideoCLIP†	Cap. Avg.	74.5	94.5	97.9
VideoCLIP†	DTW	56.0	89.9	96.3
TempCLR(Ours)	Cap. Avg.	74.5	94.6	97.0
TempCLR(Ours)	DTW	**83.5**	97.2	**99.3**
TempCLR(Ours)	OTAM	83.4	**97.3**	**99.3**

*:reported in Chen et al. (2021), †: our implementation

Table 4: Comparison on action recognition

Approach	Accuracy
TSN++*	33.6
CMN++*	34.4
RTRN++*	38.6
OTAM(Cao et al., 2020)	42.8
RTX(Perrett et al., 2021)	42.0
MTFAN(Wu et al., 2022)	**45.4**
Baseline(ResNet-50)	37.2
TempCLR (ours)	44.9

*:Results are reported in Cao et al. (2020)

One-shot action recognition is summarized in Table 4. For Baseline(R_7), it use the embeddings extracted by ResNet50 and use OTAM as measurement for action recognition. Then, our TempCLR differs OTAM by using applying self-supervised learning and generating negative sequences though shuffling positive sequence, while Cao et al. (2020) applies meta-training (Snell et al., 2017) and uses action labels. RTX also employs a Transformer but only learns the temporal order between a pair/triple of frames and the learning efficiency is limited. However, our self-supervised learning approach is more effective and is more close to the SOTA.

5 Discussion

5.1 Negative Sample Selection on Video Action Recognition

Our approach follows the contrastive learning framework and minimizes the InfoNCE (Oord et al., 2018) loss. For each anchor sequence $\mathbf{S}_a$, we generate negative samples $\mathbf{S}_n$ from the positive sample $\mathbf{S}_p$ by shuffling the order and breaking the order consistency within $(\mathbf{S}_a, \mathbf{S}_p)$. Then, we discuss relevant alternatives for the regularization and summarize the results on Table 5.

Table 5: Negative sampling strategy

Negative Strategy	Recall
un-paired	48.0
all-unit	49.3
within-seg	46.4
seg-only	52.1
seg-unit	**52.5**

Table 6: Full-video retrieval (w/ background)

Approach	Measure	R@1	R@5	R@10
VideoCLIP	DTW	56.7	93.1	98.9
VideoCLIP	OTAM	53.9	91.3	97.0
TempCLR	DTW	70.4	93.8	97.9
TempCLR	OTAM	72.2	94.5	97.7
TempCLR w/ OTAM	DTW	66.5	93.1	96.6

On Video-Paragraph pretraining, for each paragraph (anchor), besides shuffle the segments first and then shuffle the clip embeddings within each segment (*seg-unit*), we can also only shuffling the segments while maintaining the clip order within each segment (*seg-only*). Meanwhile, an intuitive

strategy is to follow Chen et al. (2020a) and use the video which is unpaired with the anchor to build negative samples (*unpaired*). In addition, we can also directly shuffle all clips embeddings in $\mathbf{S}_p$ (*all-unit*), or just keep the segment order and only shuffle the clip embeddings within each segment (*within-seg*)

As shown in Table 5, we compare the performance on CrossTask under *Supervised*. Since the distance between video and unpaired paragraph has already been high and can be easily distinguished, *unpaired* does not clearly improve the performance. According to R_{3-5}, since each caption is supposed to align with all of the clips in the paired segment, breaking the order of segments in $\mathbf{S}_a$ is essential for modelling the temporal dynamics. When $\mathbf{S}_a$ is compared with $\mathbf{S}_n$ under DTW, we assume the derived optimal matching can be used to indicate the most confusing case regarding the clip-caption matching. In this way, by minimizing $\mathcal{L}_{seq}$, the network is then trained to distinguish the confusing cases between clips due to high visual similarity. When the segment order is preserved instead, comparing with VideoCLIP baseline (R_4 in Table 1 (right)), purely shuffling the clip order within each segment does not help and may even lead to less the generalization ability. In contrast, when the segment order is broken, shuffling the clip order within each segment further can serve as data augmentation, which can improve the test performance slightly from 52.1 to 52.5. When all units are directly shuffled, the performance can be improved slightly since the clips of the same segment may have been sparsely distributed in the full sequence, and it is too hard for the model to learn.

Meanwhile, instead of setting the paragraph as anchor, we can also build positive and negative pairs for each sequence draw from video (*visual-anchor*). However, shuffling the text embeddings w.r.t. a video sequence is equivalent to (seg-only). As such, the performance by *visual-anchor* is also high and around 52.0.

For the matching between videos, as each video instance is assigned with an action class label but no segment annotation is provided, for each anchor video, besides shuffling all frame embeddings in $\mathbf{S}_p$ (*w/o label*, used in our approach), the alternative negative generation strategy it to select video samples from different categories (*w/ label*). However, as shown in Table 8, since the distance between different video instances has already been high, the performance gain is not significant.

5.2 Component Analysis and Ablation Study

Table 7: Caption-Clip Match

Approach	Correct Match
VideoCLIP	79.6%
TempCLR	91.2%

Table 8: Ablation study on CrossTask (Left) and Sth-Sth V2 (Right)

PT \ DS	w/o $\mathcal{L}_{seq}$	w/ $\mathcal{L}_{seq}$
VideoClip	47.3	52.5
TempCLR	49.3	52.0

PT: Pre-train, DS: Downstream

Approach	Diff	Same
DTW	37.5	44.5
OTAM	37.3	44.9

Optimal matching through DTW. When we directly measure the global distance between video and paragraph, it is also very important to ensure the matched units are also semantically close to each other. As shown in Table 7, we consider the full-video retrieval and check the number of correctly matched clip-caption pairs. Our TempCLR can both improve the recall in video retrieval than baseline VideoCLIP, and correctly match more caption-clip pair in the optimal matching derived from DTW. As such,

Attention in Transformer with TempCLR. In summary, our approach TempCLR is capable to provide consistent performance gain over the three task types under six different setups. attention embeds temporal background into the each single embedding, but not enough The attention mechanism in transformer architecture has been widely applied to model the correlation for long sequences, *i.e.*, long videos and paragraphs. However, purely using the attention mechanism cannot always ensure finding an optimal solution. For the attention mechanism, it is hard and may take a lot of training data to really make the attention mechanism to pay less attention on the background region.

For example, as the annotation regarding time period is not always precise, it is pretty common that the label in large-scale / videos data is noisy [(for To check, whether the noisy label will cause significant performance difference)] By training a transformer structure directly without annotation, the performance is not quite high.

For the full-video retrieval, we gather the aligned sentence-clip embeddings. As shown in Fig. **??**, our method not only can better retrie the videos and more sentences can be correctly aligned with the corresponding vide clips. As such, it can also provide proof regarding the assumption that the matched units have learned global information from the network.

Also, when the annotation is not available, according to the visualization from OTAM, the matched videos are also closed to each other given the action background (the visualization in OTAM Cao et al. (2020)). As such, using OTAM or DTW can provide a weakly supervised signal for the case where no or noisy label is provided.

(We also notice) Comparison with CRM (Huang et al., 2021b), theirs only compare the pairs. However, ours generalize to multi-instance under InfoNCE loss form and specifically focus on the hard negative resulted from the visual similarity to better distinguish the spatial and temporal.

(We also notice) the AlignNet Wang et al. (2020) has been proposed to provide dense alignment between continuous video and audio signals. It focus on the multi-scale (for diverse distortion and speed change) and use layers of different level to wrap and integrade the bottom layer with current layer, which is then for cross-modal analysis. Our approach is for sparse alignment and matching. but orthognal w.r.t. each other.

***Supervised* on CrossTask**. For action step localization, as summarized in Table 8(Left), we study the effect of $\mathcal{L}_{seq}$ in pre-training (PT) stage and downstream finetuning (DS) stage on CrossTask train set. For TempCLR, *i.e.*, applying $\mathcal{L}_{seq}$ is in PT stage, as the model has been trained to model global temporal order of the full sequence, the performance by finetuning on CrossTask train set can also be improved. Then, finetuning on CrossTask train set from TempCLR can underperform slightly than finetuning from VideoCLIP. We think the reason is because the model may overfit to HT100M as we only use 2.7% of training set for in PT. However, on both VideoCLIP and TempCLR, applying $\mathcal{L}_{seq}$ in DS can improve the performance significantly.

DTW Vs. OTAM. For self-supervision for video matching, In addition to shuffling the embeddings in the synthesized $\mathbf{S}_a$, we can also shuffle the embeddings of other video instances as $\mathbf{S}_a$. As shown in Table 8(Right), setting negative sequences by shuffling the embeddings from other videos does not clearly help the few-shot accuracy. As mentioned in Sec. 3.3, DTW has a strict boundary assumption. However, as the frame sampling process can be random, the first frame in the sampled sequence can be background or unrelated with the action. As such, using OTAM in our TempCLR does help improve the performance slightly. However, during pretraining on HowTo100M, the way to build positive samples has already guarantee the assumption that $M(1,1) = 1$ and $M(N_a, N_p) = 1$. As such, the performance by OTAM and DTW is similar ($R_{7,8}$ in Table 3, more results can be found in appendix).

6 Conclusion

Representation learning has been successful in short video clips and has shown promising transfer ability in downstream tasks. However, the temporal modelling for long video is still challenging due to the temporal dynamic/dependence over a long-time span. Thus, a long video is typically represented as a sequence of clip features and compared with other sequences. [spatio can dominate, similar visual content as shortpath, hard to disentangle temporal from spatial] In this paper, we focus on the design of representation learning strategy which is friendly to align the feature sequence extracted from the long video with other sequences with close semantic meaning. We propose a temporal alignment (Tar) loss upon contrastive learning. For every two sequence pair, we first calculates the pairwise similarity between units in two sequences and then use the maximum matching similarity through dynamic temporal wrapping as the similarity of the sequence pair. Then, to generate negative samples, we break the order within each sequence by shuffling the units. By minimizing the InfoNCE loss within the contrastive learning, the model is trained to [XXXX]. As the semantic meaning can be conveyed from different modalities, *i.e.*, vision-language, we evaluate our method on action step localization and multi-sentence video retrieval. We also reformulate the few-shot action recognition as sequence alignment task to measure the sequence matching on vision-only datasets. Experimental study has shown performance gain on all three tasks on five dataset. Detailed ablation studies are provided to justify the selection of each component.

REFERENCES

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4575–4583, 2016.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2022.

Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10618–10627, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8012–8021, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020c.

Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3884–3892, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Avinash K Dixit, John JF Sherrerd, et al. *Optimization in economic theory*. Oxford University Press on Demand, 1990.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3299–3309, 2021.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pp. 214–229. Springer, 2020.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020.

Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pp. 4086–4093, 2015.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7199–7208, October 2021a.

Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7199–7208, 2021b.

Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *European Conference on Computer Vision*, pp. 425–442. Springer, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5016–5025, 2022.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

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. *arXiv preprint arXiv:2002.06353*, 2020.

Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10573–10582, 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2630–2640, 2019.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889, 2020.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pp. 69–84, 2007.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational cross transformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 475–484, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1255–1265, 2021.

Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE transactions on Multimedia*, 9(7):1396–1403, 2007.

Yuhan Shen, Lu Wang, and Ehsan Elhamifar. Learning to segment actions from visual and language instructions via differentiable weak sequence alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10156–10165, 2021.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20020–20029, 2022.

Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4): 35–43, 2001.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–7473, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianren Wang, Zhaoyuan Fang, and Hang Zhao. Alignnet: A unifying approach to audio-visual alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3309–3317, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9151–9160, 2022.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 305–321, 2018.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800, 2021.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19163–19173, 2022.

Jiashuo Yu, Junfu Pu, Ying Cheng, Rui Feng, and Ying Shan. Self-supervised learning of music-dance representation through explicit-implicit rhythm synchronization. *arXiv preprint arXiv:2207.03190*, 2022.

Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *European Conference on Computer Vision*, pp. 525–542. Springer, 2020.

Feng Zhou and Fernando Torre. Canonical time warping for alignment of human behavior. *Advances in neural information processing systems*, 22, 2009.

Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 751–766, 2018.

Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8746–8755, 2020.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3537–3545, 2019.

A Appendix

You may include other additional sections here.