Bi-calibration Networks for Weakly-Supervised Video Representation Learning

Fuchen Long · Ting Yao · Zhaofan Qiu · Xinmei Tian · Jiebo Luo · Tao Mei

Received: 1 August 2022 / Accepted: 6 March 2023 / Published online: 31 March 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

The leverage of large volumes of web videos paired with the query (short phrase for searching the video) or surrounding text (long textual description, e.g., video title) offers an economic and extensible alternative to supervised video representation learning. Nevertheless, modeling such weakly visual-textual connection is not trivial due to query polysemy (i.e., many possible meanings for a query) and text isomorphism (i.e., same syntactic structure of different text). In this paper, we introduce a new design of mutual calibration between query and text to achieve more reliable visual-textual supervision for video representation learning. Specifically, we present Bi-Calibration Networks (BCN) that novelly couples two calibrations to learn the correction from text to query and vice versa. Technically, BCN executes clustering on all the titles of the videos searched by an identical query and takes the centroid of each cluster as a text prototype. All the queries constitute the query set. The representation learning of BCN is then formulated as video classification over text prototypes and queries, with text-to-query and query-to-text calibrations. A selection scheme is also devised to balance the two calibrations. Two large-scale web video datasets paired with query and title, named YOVO-3M and YOVO-10M, are newly collected for weakly-supervised video feature learning. The video features of BCN with ResNet backbone learnt on YOVO-3M (3M YouTube videos) obtain superior results under linear protocol on action recognition. More remarkably, BCN trained on the larger set of YOVO-10M (10M YouTube videos) with further fine-tuning leads to 1.3% gain in top-1 accuracy on Kinetics-400 dataset over the state-of-the-art TAda2D method with ImageNet pre-training. Source code and datasets are available at https://github.com/FuchenUSTC/BCN.

Keywords
Weakly-supervised video representation learning · Action recognition · Large-scale web video dataset

1 Introduction

With the rise of deep learning technologies, there has been a steady momentum of breakthroughs on video representation learning (Ji et al., 2012; Wang et al., 2013; Simonyan & Zisserman, 2014; Carreira & Zisserman, 2017; Li et al., 2020b; Yang et al., 2020b; Zhao et al., 2021; Liu et al., 2022; Mettes et al., 2021; Wu et al., 2021; Kong & Fu, 2022; Long et al., 2022a). The achievements rely heavily on the requirement to have large amount of labeled data for fully-supervised training. In practice, acquiring the annotations of videos is very expensive and time-consuming. Therefore, recent research (Ghadiyaram et al., 2019; Li & Wang, 2020) study the alternative regime of web data, which is largely available and freely accessible by search engines, for weakly-supervised learning.
learning. These approaches usually treat the weakly visual-textual connection as a reliable signal and directly maximize the similarity between them or alleviate the challenge of noise (e.g., incorrect or irrelevant text) in web data for training, but seldom explore the inherent property of videos or texts. For example, is it reasonable that the representations of all the searched videos from an identical query cluster around a single prototype? or if the surrounding texts of two videos are close in representation space, should the representations of the two videos also be in proximity?

In order to answer the two questions, let us look at the two examples illustrated in Fig. 1. The upper one showcases two returned videos when searching for the query of “zebra” in a text-based video search engine. The query “zebra” de facto contains two types of search intention: an animal or a brand of printer, and the videos in response to the two meanings are quite distinct in visual appearance. The phenomenon is known as “query polysemy.” In this case, taking the two families of videos as one class will inevitably mislead video representation learning through either classification or cross-view embedding. We propose to mitigate the issue by leveraging the calibration from surrounding text (long textual description, e.g., video title in this work) of the videos. The titles of “why don’t we ride zebras? Episode 1” and “Zebra iMZ320 Product overview | iOS mobile printer” provide rich information about the video content and are more descriptive. The videos are naturally grouped into two clusters based on valuable supervision of titles and each cluster corresponds to one semantic meaning, potentially making video representation learning more discriminative. In contrast, robust learning of video representation also necessitates using the query to regulate the text. Taking the lower case as an example, the titles of “What’s inside a Baseball?” and “What’s inside a Soccer Ball?” are in the same syntactic structure but the objectives are different. We define this as “text isomorphism.” The textual representation will be close to each other since the isomorphism. Solely capitalizing on the title features will tend to draw the two videos close in representation space but unfortunately the videos describe two different objects, which are indicated by search queries. As a result, query information here is a rewarding signal to adjust visual-textual correlation and further improve video representation learning.

By delving into mutual calibration across query and text for weakly-supervised video representation learning, we present a novel Bi-Calibration Networks (BCN) architecture with the backbone of 3D ConvNet. Specifically, we employ the off-the-shelf BERT (Devlin et al., 2018) model to extract the title features and cluster all the titles of the videos searched by an identical query into different clusters. The centroid of each cluster is then taken as a text prototype and put into the text prototype set. Primary text supervision is measured as the cosine similarity between video title and all text prototypes, which regulates the video-to-text (v2t) projection (embedding) for video classification over text prototypes. Similarly, we take the “one-hot” vector built directly on the queries as primary query supervision to optimize video-to-query (v2q) projection (embedding) for video classification over queries. Next, v2t/v2q projection starts the text-to-query (t2q) or query-to-text (q2t) calibration procedure, in which BCN aggregates/decomposes the predictions on cluster/query level in a bottom-up/top-down way to produce the t2q/q2t correction. The two corrections refine the primary query or text supervision to further optimize v2q/v2t projection. Moreover, we devise a selection scheme to balance the two calibrations. The whole architecture is optimized by minimizing query and text classification loss.

The main contribution of this work is the proposal of exploring the cross correction between query and text to boost weakly-supervised video representation learning. This also leads to a better view of why query and text could complement to each other to validate visual-textual connections, and how to integrate the correction across the two into video representation learning framework. Two large-scale web video datasets are proposed for the weakly-supervised video representation learning and extensive experiments on the datasets demonstrate the effectiveness of our BCN framework.

2 Related Work

We briefly group the related works into three categories: supervised, unsupervised and weakly-supervised video representation learning with respect to the utilization of clean labels, no label or noisy labels for model training.

The early works (Simonyan & Zisserman, 2014; Karpathy et al., 2014; Ng et al., 2015; Wang et al., 2016; Feichtenhofer et al., 2016; Diba et al., 2017) of supervised video representation learning are extended from image representation by applying 2D CNN on video frames. For instance, Karpathy et al. (Karpathy et al., 2014) leverage spatio-temporal convo-
tions for representation learning on the stacked frame-level features. To further capture the motion information, the well-known two-stream architecture (Simonyan & Zisserman, 2014) and its variants (Wang et al., 2016; Diba et al., 2017; Wang et al., 2018) are devised by executing 2D CNN on optical flow. Though the methods improve the representation learning by formulating motion pattern, performing 2D CNN on video frames still limits the capacity of modeling long-range temporal dynamics. To alleviate this problem, LSTM-RNN (Ng et al., 2015) captures the long-term dependencies in videos by utilizing a long-short term memory (LSTM) auto-encoder. To treat the video clip as a temporal evolution unit rather than a sequence of independent frames, 3D CNN structures (Tran et al., 2015; Qiu et al., 2017; Carreira & Zisserman, 2017; Harra et al., 2018; Tran et al., 2018; Long et al., 2019, 2020; Qiu et al., 2021) are proposed to boost video feature learning on large-scale supervised video datasets (Carreira & Zisserman, 2017; Ghanem et al., 2018). More recently, the video transformer works (Bertasius et al., 2021; Fan et al., 2021; Arnab et al., 2021; Long et al., 2022b) are extended from the vision transformers (Dosovitskiy et al., 2021; Liu et al., 2021; Li et al., 2022; Yao et al., 2022a, b) in image domain to learn video representation. For instance, TimeSformer (Bertasius et al., 2021) explores five different variants of space-time attention and MViT (Fan et al., 2021) provides an alternative to formulate video Transformer in a multi-scale manner. We choose 3D ConvNet as the backbone for BCN training to achieve good balance between performance and efficiency.

Unsupervised video representation learning is one kind of technique to employ unlabeled data for video feature learning. The related works leverage various supervision from the video data itself to build pretext tasks, such as frame/clip order prediction (Misra et al., 2016; Fernando et al., 2017; Wei et al., 2018; Xu et al., 2019), motion estimation (Agrawal et al., 2015; Pathak et al., 2017), temporal cycle-consistency learning (Dwibedi et al., 2019; Wang et al., 2019), temporal coherence learning (Mobahi et al., 2009; Wang & Gupta, 2015), pixel-level displacement prediction (Liu et al., 2017; Wang et al., 2019), frame reconstruction (Srivastava et al., 2015; Finn et al., 2016; Luo et al., 2017), contrastive learning (Feichtenhofer et al., 2021; Cai et al., 2020; Li et al., 2021b; Yao et al., 2021; Li et al., 2021c) and HOG feature prediction (Wei et al., 2021). To further improve the descriptive ability of video representation, weakly-supervised methods (Ghadiyaram et al., 2019; Li & Wang, 2020; Miech et al., 2020) focus on utilizing the weak supervision from web video data (Miech et al., 2019; Pan et al., 2022). For example, Ghadiyaram et al. (Ghadiyaram et al., 2019) explore the influences of different aspects in tags (e.g., label space) for weakly-supervised learning. Furthermore, to mine the knowledge from title of web videos, CPD (Li & Wang, 2020) learns the video representation by making visual-textual pair close to each other. More recently, by adapting the webly-supervised pre-trained image model CLIP (Radford et al., 2021) in video domain, Wang et al. (Wang et al., 2021b) obtains good performances on downstream action recognition task via promote learning. Nevertheless, these weakly-supervised video representation learning methods are still facing the challenge of query polysemy or text isomorphism when directly exploiting query or text as supervision. Even some traditional works (Berg et al., 2006; Schroff et al., 2007; Saenko & Durrell, 2008) handled query polysemy in web data collection, our proposal alleviates it during training stage without any complicated priors.

In short, our approach belongs to weakly-supervised video representation learning techniques. Different from the aforementioned methods which solely rely on the supervision of query or text, our BCN in this paper contributes by studying not only how to explore supervisory signal from query and text simultaneously, but also how mutual calibration between query and text information could be leveraged to enhance weakly-supervised video representation learning.

3 Bi-calibration Networks

In this section, we introduce the Bi-Calibration Networks (BCN) that performs mutual calibration between query and text to facilitate weakly-supervised video representation learning. Specifically, BCN formulates the problem as the video classification over the query and text labels. To better understand the spirit of our BCN design, we first introduce the problem formulation of the unique pretext task in BCN for weakly-supervised video representation learning. After that, the detailed BCN architecture is further elaborated.

3.1 Problem Formulation

The weakly-supervised video representation learning on web data is usually formulated as the task of training a visual encoder based on the weak annotations, e.g., searched queries or surrounding texts (e.g., titles). Nevertheless, solely relying on the query or text supervision may face the problem of query polysemy and text isomorphism as mentioned before. The visual-textual connection is brittle when the same query has different meanings or the different surrounding texts share the same syntactic structure. To alleviate these issues, we capitalize on both of the query and text supervision to pre-train the visual encoder, and explore the mutual calibration across them to achieve more reliable visual-textual connection for video feature learning. We first formulate the pretext task of video feature learning as the classification over query and text. For video classification over query, we follow the natural solution (Ghadiyaram et al., 2019) that treats
Fig. 2 The conceptual illustration of weakly-supervised video representation learning on a query supervision; b text supervision; c both of query and text supervision with bi-calibration in BCN.

![Diagram](image)

(a) Feature learning on query supervision
(b) Feature learning on text supervision
(c) Feature learning on query and text supervision through bi-directional calibration

Fig. 3 An overview of our Bi-Calibration Networks (BCN). In general, BCN a first extracts the video representation of input video via 3D CNN, and then feed it into two branches, i.e., video-to-query (v2q) and video-to-text (v2t) projections. The output query distribution over the whole query phrase as one category and learn to classify the video as the corresponding category (see Fig. 2a). Moreover, as depicted in Fig. 2b, we convert the video title into the soft labels and employ them for video classification over text. To further improve the quality of primary supervision and facilitate video representation learning, we design a bi-directional calibration learning paradigm to correct query and text supervision. As shown in Fig. 2c, the primary query and text labels are refined by both text-to-query and query-to-text calibration through the video-to-query and video-to-text projections, aiming to strengthen the weakly-supervised video representation learning.

### 3.2 An Overview of BCN

Figure 3 details an overview of the BCN framework and two coupled calibration modules, i.e., t2q and q2t calibration. Specifically, for each input video, BCN first utilizes a 3D CNN to extract video representation and feed it into two branches, i.e., v2q and v2t projections. Two kinds of probability distributions (i.e., the query distribution over queries and the text distribution over all text prototypes) are achieved to trigger each calibration. Note that all the titles of the videos searched by an identical query are initially grouped into multiple clusters, and each text prototype corresponds to the centroid of each cluster. In this way, we naturally obtain two kinds of primary supervision for each video, i.e., the primary query supervision (the “one-hot” vector of query) and the primary text supervision (the cosine similarity between video title and all text prototypes), that are used to optimize v2q and v2t projections (embedding), respectively. After that, BCN starts t2q/q2t calibration procedure by aggregating/decomposing the text or query distribution into t2q/q2t correction in a bottom-up/top-down fashion, respectively. The learnt t2q/q2t correction is further integrated with the primary query/text supervision, yielding the refined query or text.
text supervision to strengthen the regulation of each branch. During training, a selection scheme is utilized to balance the two calibration procedures.

3.3 Video-to-Query/Text Projection Branch

The ultimate target of BCN is to train the 3D CNN backbone for video representation learning through two pretext tasks of query and text classification. Therefore, based on the extracted video feature, we involve two parallel v2q and v2t projection branches to perform the two pretext tasks, which can be optimized with the corresponding query and text supervision.

Primary Query and Text Supervision

Given all the training web videos paired with the searched queries and surrounding text (i.e., titles), one natural way to represent each query or text as query/text supervision is to directly construct query/text label set for classification. Therefore, for each query phrase (with one or multiple words), we treat it as one category label and take its “one-hot” vector \( y^q \in \mathbb{R}^K \) in query label set (query number: \( K \)) as the primary query supervision. However, the robustness of this recipe is brittle when applying it to represent surrounding texts, since the inherently semantic distribution among different titles is unexploited. Inspired by the classical bag-of-words paradigm for feature representation (Lazebnik et al., 2006), we leverage such paradigm to construct text label set by delving into the diverse semantic structure of titles. Formally, we first extract the text feature \( f^t \) of each video title by the off-the-shelf language model BERT (Devlin et al., 2018). Next, for the \( k \)-th query, we perform clustering on all the titles of the videos searched by this query through k-means algorithm, leading to \( m_k \) clusters. The cluster number \( m_k \) is automatically set by a statistic-based clustering estimation method (Gap Statistic (Tibshirani et al., 2001)). The centroid \( B_i \) of the \( i \)-th cluster is then defined as one text prototype, which is computed as the average of all text features of video titles in that cluster. Accordingly, we construct the text label set based on all the \( M = \sum_{k=0}^{K-1} m_k \) text prototypes derived from \( K \) queries, which naturally characterize the underlying semantic structure of titles. Based on this text label set, we interpret the primary text supervision \( y^t = [y^t_0, y^t_1, \ldots, y^t_{M-1}] \in \mathbb{R}^M \) of each surrounding text \( f^t \) as the cosine similarities between \( f^t \) and all the \( M \) text prototypes \( \{B_i \mid i = 0, 1, \ldots, M-1\} \):

\[
y^t = \text{softmax}(\cos(f^t, B_i)) \mid i = 0, 1, \ldots, M-1, \tag{1}
\]

where \( \text{softmax}(\cdot) \) is the softmax with temperature parameter \( 1/M \) and \( \cos(\cdot, \cdot) \) represents cosine similarity.

One case of the clustering for video titles searched by the query of “swimming breast stroke” is visualized in Fig. 4. There are two meaningful clusters as shown in the figure. The first cluster has an emphasis on the daily breaststroke (e.g., tutorial) while most of the titles in the second cluster tend to describe the championship of swimming. The differences in between are reflected in visual content. Thus, the text supervision based on the similarity between title feature and each prototype well captures the semantic relation.

Video-to-Query Projection

The v2q projection branch is especially designed to perform the pretext task of query classification. Concretely, depending on the input video feature \( f^v \) extracted by 3D CNN, the v2q projection branch infers its query distribution \( p^q = \{p^q_0, p^q_1, \ldots, p^q_{K-1}\} \) over all queries through a fully-connected layer. Note that \( p^q_k \) is the estimated probability of assigning this video to the \( k \)-th query. The v2q projection branch can be directly optimized with the primary query supervision \( y^q \), and we measure the query classification loss as the softmax cross-entropy loss as follows:

\[
L^q = - \sum_{k=0}^{K-1} I_{k=y^q} \log(p^q_k). \tag{2}
\]

where the indicator function \( I_{k=y^q} = 1 \) if the label value of \( k \)-th query in \( y^q \) is 1, otherwise \( I_{k=y^q} = 0 \).

Video-to-Text Projection

In analogy to the v2q projection branch, we design another v2t projection branch to conduct the pretext task of text classification. In particular, the v2t projection branch takes the video feature \( f^v \) as the input, and learns to estimate the text distribution \( p^t = \{p^t_0, p^t_1, \ldots, p^t_{M-1}\} \) over all text prototypes. \( p^t_i \) denotes the probability of the video belonging to the \( i \)-th cluster. Accordingly, we optimize the v2t projection branch with the primary text supervision \( y^t = [y^t_0, y^t_1, \ldots, y^t_{M-1}] \) and the text classification loss for this proxy task is calculated as the softmax cross-entropy loss:

\[
L^t = - \sum_{i=0}^{M-1} y^t_i \log(p^t_i). \tag{3}
\]
3.4 Text-to-Query Calibration

The most typical way to optimize v2q projection branch is to use the primary query supervision for query classification as in Eq.(2). However, this oversimplifies the proxy task by assuming that all videos searched by an identical query belong to one class, while ignoring the phenomenon of query polysemy (i.e., the coexistence of many possible meanings for a query). That will inevitably mislead video representation learning. To alleviate this problem, we devise a text-to-query (t2q) calibration module that further regulates the v2q projection with additional calibration from video surrounding texts (i.e., titles).

**Text-to-Query Correction** The t2q calibration module first transforms the text distribution $p^t$ derived from the v2t projection branch into the t2q correction $\hat{p}^q \in \mathbb{R}^K$ by aggregating all the probabilities of text prototypes belonging to the same query in a bottom-up manner as shown in Fig. 5a:

$$
\hat{p}^q_k = \sum_{i \in I_k} p^t_i, \ s.t. \ k \in \{0, 1, \ldots, K - 1\},
$$

where $\hat{p}^q_k$ denotes the $k$-th correction value in t2q correction $\hat{p}^q$ and represents the aggregated probability with regard to the $k$-th query in the query set. $I_k$ is the index set of the text prototypes for the $k$-th query.

**Refined Query Supervision** Next, the t2q correction $\hat{p}^q$ serves as the additional calibration from the text distribution over all text prototypes, aiming to refine the primary query supervision $y^q$ with more semantic meanings derived from v2t projection branch. Inspired by the idea of self-contained confidence (SCC) (Yang et al., 2020c) for image webly-supervised learning, we take the estimated detached query distribution $p^q$ as the query SCC and further utilize it to balance the primary query supervision and t2q correction. As illustrated in Fig. 5a, the refined query supervision $R^q$ is computed as follows:

$$
R^q = \hat{p}^q \circ (1 - p^q) + y^q \circ p^q,
$$

where the operation $\circ$ denotes the element-wise multiplication and $1 \in \mathbb{R}^K$ is the vector of all ones. The underlying assumption behind Eq.(5) is that if the self-contained confidence of the query is higher, the refined query supervision will prefers to trust primary query supervision. Otherwise, the refined query supervision tends to be heavily influenced with the t2q correction. Finally, the t2q calibration module leverages the refined query supervision $R^q$ to further optimize the v2q projection branch, and the query classification loss in Eq.(2) is thus reformulated as:

$$
\hat{L}^q = - \sum_{k=0}^{K-1} R^q_k \log(p^q_k),
$$

3.5 Query-to-Text Calibration

Recall that the aforementioned optimization of the v2t projection branch (see Eq.(3)) solely hinges on the primary text supervision for text classification proxy task. Nevertheless, in the case of text isomorphism (i.e., titles share the same syntactic structure but refer to different semantics), the semantic discriminativeness of video representations learnt in this way may be easily overwhelmed since the similar syntactic structure leads to closed textual representation. To address the issue, a query-to-text (q2t) calibration module is designed to guide the optimization of v2t projection with the additional high-level semantic supervision from query.

**Query-to-Text Correction** In the q2t calibration module (Fig.5b), we first calculate the q2t correction $\hat{p}^t \in \mathbb{R}^M$ by evenly decomposing each element (e.g., the probability $p^q_k$ of $k$-th query) of query distribution to the correction values over all the text prototypes belonging to that query:

$$
\hat{p}^t_i = \frac{1}{m_k} p^q_k, \ s.t. \ i \in I_k, \ k \in \{0, 1, \ldots, K - 1\},
$$

where $\hat{p}^t_i$ denotes the correction value of the $i$-th text prototype in q2t correction $\hat{p}^t$ and $m_k$ is the number of text prototypes belonging to the $k$-th query.

**Refined Text Supervision** After that, we improve the primary text supervision with the q2t correction, leading to the refined text supervision to further regulate the v2t projection branch. Similar to the formulation of refined query supervision, we treat the estimated text distribution $p^t$ as the text self-contained confidence (SCC). As shown in Fig. 5b, the refined text supervision $R^t$ is calculated by fusing q2t correction $\hat{p}^t$ and the primary text supervision $y^t$ according to the text self-contained confidence $p^t$:

$$
R^t = \hat{p}^t \circ (1 - p^t) + y^t \circ p^t.
$$

Based on the refined text supervision $R^t$, the q2t calibration module further optimizes the v2t projection branch, and the
text classification loss in Eq.(3) is thus reformulated as:
\[
\hat{L}' = - \sum_{i=0}^{M-1} R'_i \log(p'_i), 
\]
where \(R'_i\) is the new ground-truth of the \(i\)-th text prototype in the refined text supervision \(R^t\).

### 3.6 Network Optimization

**Algorithm 1: Calibration Selection Scheme**

| Input       | Model \(M\) of the first training stage, web videos with query and title; two thresholds \(\varepsilon_q\) and \(\varepsilon_t\); |
|-------------|--------------------------------------------------|
| Output      | Output model \(\hat{M}\) of BCN; |
| 1 \(\text{while} \ n \leq N - 1 \) \(\text{do}\) | |
| 2 \(\text{Network forward to obtain probabilities } p^q\) and \(p^t\); | |
| 3 \(\text{Compute } y^q\) and \(y^t\) according to Eq.(4) and Eq.(7); | |
| 4 \(\text{Compute } y^q\) and \(y^t\) by Eq.(5) and Eq.(8); | |
| 5 \(\text{if } \|p^q \circ y^q\|_2 < \varepsilon_q \text{ and } \|p^t \circ y^t\|_2 > \varepsilon_t \text{ then}\) | |
| 6 \(\text{Optimize } \hat{M} \text{ by } \hat{L}^q \text{ and } L^t; (q2t calibration)\) | |
| 7 \(\text{else if } \|p^q \circ y^q\|_2 > \varepsilon_q \text{ and } \|p^t \circ y^t\|_2 < \varepsilon_t \text{ then}\) | |
| 8 \(\text{Optimize } \hat{M} \text{ by } \hat{L}^q \text{ and } L^t; (t2q calibration)\) | |
| 9 \(\text{else if } \|p^q \circ y^q\|_2 > \varepsilon_q \text{ and } \|p^t \circ y^t\|_2 < \varepsilon_t \text{ then}\) | |
| 10 \(\text{Optimize } \hat{M} \text{ by } \hat{L}^q \text{ and } L^t; (q2t calibration)\) | |
| 11 \(\text{end while}\) | |
| 12 \(\text{return } \hat{M}\) | |

During training, we adopt a two-stage strategy to optimize the whole architecture of our BCN. In the first stage, we optimize the BCN framework with the typical query and text classification losses \((L^q\) in Eq.(2) and \(L^t\) in Eq.(3)) simultaneously, irrespective of any calibration modules. Thus, we can obtain the query and text self-contained confidence as illustrated by Yang et al. (2020c). In the second stage, the BCN framework is further fine-tuned with two coupled calibration modules. Here, we design a selection scheme to balance the two calibration modules according to self-contained confidence weighted by the primary query and text supervision \((p^q \circ y^q\) and \(p^t \circ y^t\)) in each module.

More specifically, if \(\|p^q \circ y^q\|_2 < \varepsilon_q\) and \(\|p^t \circ y^t\|_2 > \varepsilon_t\) \((\varepsilon_q\) and \(\varepsilon_t\) are two thresholds), this case implies that the primary text supervision \((y^t)\) is more reliable than the primary query supervision \((y^q)\). Therefore we select the t2q calibration module, and the BCN framework is optimized with the modified query classification loss \(\hat{L}^q\) in Eq.(6) plus the typical text classification loss \(L^t\) in Eq.(3). In contrast, if \(\|p^q \circ y^q\|_2 > \varepsilon_q\) and \(\|p^t \circ y^t\|_2 < \varepsilon_t\), the primary query supervision is supposed to be more reliable than the primary text supervision, and the q2t calibration module is selected for optimization. In that case, the objective of BCN framework consists of the modified text classification loss \(\hat{L}^t\) in Eq.(9) and the typical query classification loss \(L^q\) in Eq.(2).

Otherwise, we optimize BCN with the typical query and text classification losses as in the first stage. The weight for each loss is set as 1.0 empirically.

Algorithm 1 details the processing of the selection scheme of the two calibrations in our BCN.

### 4 Experiments

The experiments of weakly-supervised video representation learning are conducted on the newly-created YouTube video (YOuTube VideoO) datasets, namely YOVO-3M and YOVO-10M, particularly paired with query and title. We then empirically verify the merit of BCN on three scene-related action recognition datasets: Kinetics-400 (Carreira & Zisserman, 2017), UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011), and two interaction-related action recognition datasets: Something-Something V1 and V2 (Goyal et al., 2017).

### 4.1 Datasets

**YOVO-3M/10M Datasets** We collect the YOVO-3M and YOVO-10M datasets characterized by the unique properties including large-scale web video data with query and title information, as well as the comprehensive and diverse video content for weakly-supervised video representation learning. Figure 6 depicts the construction pipeline of YOVO-3M and YOVO-10M datasets, which consists of six main steps, i.e., query set collection, query deduplication, query filtering, video collection on YouTube, video filtering and clip deduplication. To crawl the web videos with comprehensive visual content, we first collect all the labels from Kinetics-400 (Carreira & Zisserman, 2017), Kinetics-700 (Carreira et al., 2019), ImageNet (Deng et al., 2009) and Moments (Monfort et al., 2019) datasets as search queries. All the queries will be parsed for query deduplication. In detail, we first exploit the BERT model to extract the feature of each query phrase and then calculate the cosine similarity matrix. If the similarity of two query phrases is higher than a threshold, the two query phrases will be merged as one for the subsequent processing. After query deduplication, the number of remaining queries is 2,015. Since all queries are collected from labels of public datasets, there is no sensitive word among them. We issue each query to YouTube and about 489 videos with the titles are downloaded successfully on average. We then uniformly sample three 10-second clips from each video to build YOVO-3M dataset. To further enlarge the volume of web videos, a 3K verb vocabulary and a 13K noun vocabulary extracted from Oxford English Dictionary are utilized as additional queries. Given these extra queries, we employ the Profanityfilter toolbox with a blacklist of the sensitive words to filter out the inappropriate queries. The remaining queries
Table 1  The comparisons between YOVO-3M/10M and other video datasets from the literature

| Dataset                        | Source  | Context | Surrounding text | # of Video | # of Clip | # of Category/query | Duration (h) |
|-------------------------------|---------|---------|------------------|------------|----------|---------------------|--------------|
| UCF101 (Soomro et al., 2012)  | Labeled | Action  | None             | 13,320     | 101      | 26.7                |              |
| EPIC-KITCHENS (Damen et al., 2018) | Labeled | Action  | None             | 39,596     | 149      | 55.0                |              |
| Something-Something V1 (Goyal et al., 2017) | Labeled | Action  | None             | 108,499    | 174      | 120.4               |              |
| Something-Something V2 (Goyal et al., 2017) | Labeled | Action  | None             | 220,847    | 174      | 245.4               |              |
| Kinetics-400 (Carreira & Zisserman, 2017) | Labeled | Action  | None             | 306,245    | 400      | 850.7               |              |
| YouTube-8M (Abu-El-Haija et al., 2016) | Weakly labeled | Multi-category | None | 8,264,650 | 4800 | 500K |
| Sports-1M (Karpathy et al., 2014) | Web data | Action  | None             | 1,133,158  | 487      | 105761.4            |              |
| HowTo100M (Miech et al., 2019) | Web data | Action | Caption          | 1.22M      | 136M     | 23.6K               | 151777.8     |
| YOVO-3M                        | Web data | Multi-category | Video title | 986,031 | 2,958,092 | 2,015 | 8216.9 |
| YOVO-10M                       | Web data | Multi-category | Video title | 8,051,431 | 10,023,532 | 18,305 | 12142.1 |

are exploited to search for another set of videos with the titles. In an effort to screen out the searched videos with sensitive content, we exploit Profanityfilter and remove the videos if their titles contain sensitive words. Besides, a ResNet-50 model trained on NPDI (Avila et al., 2013) and UCF-Crime (Sultani et al., 2018) datasets is adopted to detect the adult and violent contents on the key frames of each video, aiming to remove the sensitive videos. To save the storage, we random sample only one clip for each video and combine these clips with YOVO-3M dataset to construct YOVO-10M dataset. Finally, we employ the standard clip deduplication approach (Ghadiyaram et al., 2019) to remove video clips occurring anywhere in the downstream datasets from both YOVO-3M and YOVO-10M datasets. Specifically, we extract the global feature of each frame in a video clip through Census Transform (Zabih & Woodfill, 1994) and then execute Locality Sensitive Hashing (LSH) (Gionis et al., 1999) on the feature to obtain frame-level hash codes. We then compute the hamming distance between the frame-level hash codes of each pair of frames, in which one is from a video clip from YOVO-3M/10M datasets and the other is from a video in downstream recognition datasets, e.g., Kinetics-400. We average all the distances of frame pairs across two videos as the clip-level distance in between. If the distance is lower than 2, we will remove the corresponding video clip from YOVO-3M/10M.

Table 1 summarizes the comparisons between YOVO-3M/10M and other YouTube video datasets. As shown in the table, most of the existing video datasets are collected with manual annotations for supervised video representation learning. The scale of some datasets (e.g., UCF101 and EPIC-KITCHENS (Damen et al., 2018)) is relatively small and not applicable for the large-scale video pre-training. There are also other video datasets containing huge number of video data, such as Sports-1M (Karpathy et al., 2014) and HowTo100M (Miech et al., 2019). However, the majority of the contexts in the two datasets focuses on action, thereby limiting the diversity of the video contents. Meanwhile, the surrounding text (e.g., video title) of the videos in these datasets is usually not provided. In contrast, we collect two new video datasets, i.e., YOVO-3M and YOVO-10M, with comprehensive visual contents and surrounding text for large-scale video pre-training. In detail, YOVO-3M contains 2,015 queries and 2,958,092 video clips in total, and YOVO-10M consists of 18,305 queries and 10,023,532 video clips. The two scales of the proposed datasets would be further applicable to different researchers w.r.t computational resources in the video representation learning community. Figure 7 further illustrates 32 video clips with the searched queries and the titles from YOVO-3M and YOVO-10M datasets. The showing video cases demonstrate the diverse video content in different facets, e.g., objects, sports and daily activities, for weakly-supervised video feature learning. We
have also provided the datasheets (Gebru et al., 2021) for YOVO-3M/10M on our project website.

We additionally study the relevance between the query and video content to investigate data noise on our datasets. Specifically, in YOVO-10M dataset, we random sample two sets of 1,000 video clips with object queries and action queries, respectively. We invite five evaluators to manually rate the relevance between the query and video content (0 to 4: 0: not relevant, 1: might be relevant, 2: somewhat relevant, 3: relevant, 4: very relevant). Figure 8 depicts the boxplots of the relevance rating between the videos and different kinds of queries. The average relevance rating of sampled videos derived from all queries is 2.5, which indicates the existence of noise in the searched queries for web videos. When looking at the query-video relevance across different kinds of queries, the relevance degree between videos and action queries is generally higher than that between videos and object queries. We speculate that the results may be caused by the larger complexity of objects in videos, and it is non-trivial to accurately describe a video with an object query.

**Downstream Datasets**

We evaluate our BCN on five downstream action recognition datasets. The Kinetics-400 (Carreira & Zisserman, 2017) dataset consists of 300K 10-second clips from 400 action categories. There are 20K, 40K clips in training, validation and testing sets, respectively. The UCF101 (Soomro et al., 2012) contains 13K videos from 101 action classes, and the HMDB51 (Kuehne et al., 2011) has 7K videos from 51 action categories. In UCF101 and HMDB51 datasets, there are three training/validation splits provided by the dataset organizers. Each split in UCF101 includes about 9.5K training and 3.7K validation videos, and a HMDB51 split contains 3.5K training and 1.5K validation videos. The videos in the above three datasets are mainly for scene-related action recognition. In Something-Something V1 and V2 (Goyal et al., 2017) datasets, there are about 108K and 221K

![Fig. 7 Examples of 32 video clips with searched queries and titles from YOVO-3M and YOVO-10M datasets. The video clips in each row are searched by the queries in the category list of different datasets, i.e., ImageNet, Kinetics and Moments, or in the vocabulary list of the Oxford English Dictionary](https://example.com/fig7.png)

![Fig. 8 Boxplots of user study towards the relevance between query and video content on the subset of YOVO-10M dataset. The average relevance rating is marked with the cross](https://example.com/fig8.png)
videos from 174 action categories, respectively. The training/validation/testing set consists of 86K/11.5K/11K and 169K/25K/27K videos, which are mostly for interaction-related recognition.

4.2 Experimental Settings

Implementations We parse the video titles via BERT (Devlin et al., 2018) text encoder with a comprehensive vocabulary to extract the 1,024-D features for each word token in the title. Such vocabulary list contains both English and non-English words. If the video title contains non-English words that do not appear in the list, the words will be removed. We average all the token features as text representation. The number of total clusters is 6, 819 or 65, 766 on YOVO-3M or 10M. In weakly-supervised training, we utilize the architecture of LGD-3D (Qiu et al., 2019) based on ResNet-50 (He et al., 2016) as the backbone to obtain the good trade-off between performance and training efficiency. All the parameters are trained from scratch. The dimension of the input video clips is set as 32 × 112 × 112, which is randomly cropped from the original web video. Each clip is randomly flipped along horizontal direction for data augmentation. We choose both of the two threshold \( \epsilon^q \) and \( \epsilon^t \) as 0.5 by cross validation on Kinetics-400 datasets under linear evaluation protocol. We implement BCN on Caffe (Jia et al., 2014) platform. In all the pre-training stages, the networks are trained by utilizing stochastic gradient descent (SGD) with 0.9 momentum. The initial learning rate is set to 0.08 and 0.008 in the first and second training stage, and decreased by 10% after every 200K iterations. The mini-batch size is set as 256 and the weight decay is 0.0001.

Evaluation Metrics We adopt two evaluation protocols in the downstream recognition datasets, i.e., linear model and network fine-tuning. In the former protocol, we uniformly sample 10 or 3 video clips from each video in Kinetics-400/UCF101/HMDB51 or Something-Something V1/V2 datasets, and take the 2, 048-way outputs from pool5 layer of the network backbone as the features of each clip. We average all the features of clips in one video as video representation, and a linear SVM is learnt on the training set and evaluated on validation set. In detail, the cost parameter \( c \) in SVM is set as 8.0. Both top-1 and top-5 accuracy are reported as evaluation metrics. In the latter one, we initialize the network backbone with the weakly-supervised training output model of BCN, and fine-tune/evaluate the network on the training/validation set of each dataset.

4.3 Evaluation on Primary Text Supervision

We first examine the effectiveness of primary text supervision for video representation learning, regardless of mutual calibration design. We compare the following four methods:

1. The regression method (RG) optimizes video representation through minimizing Smooth L1 loss (Girshick, 2015) between video representation and text feature.
2. Triplet Ranking algorithm (TR) learns video representation to make positive video-text pair more similar than negative pair.
3. A variant of our primary text supervision (TS) also performs clustering on video titles and builds text label set on the centroids of clusters. Each title is naturally assigned to one cluster which has the largest similarity value in \( y^i \) of Eq. (1) and represented as a binary index vector in text label set. We exploit single-label classification to regulate video representation with text supervision in TS-.
4. TS is our proposed primary text supervision in BCN, which computes cosine similarity between title and all clusters. TS can be regarded as the soft mode of TS- and accordingly tunes video feature with multi-label classification loss.

Table 2 summarizes performance comparisons of video representation learnt with different ways of primary text supervision under linear protocol on three downstream datasets. Overall, the results across different datasets consistently indicate that TS leads to a performance boost against other methods. In particular, the top-1 accuracy of our TS achieves 72.5%, 91.8% and 64.5% on Kinetics, UCF101 and HMDB51, respectively, making the absolute improvement over RG/TR by 8.0%/5.3%, 6.7%/4.6% and 6.2%/2.7%. Such results demonstrate the advantage of exploring the structure among all the titles of videos via clustering. Though both TS- and TS utilize text clustering to improve text supervision, they are different in the way that TS- represents each title as an index vector (1 for its own cluster with largest similarity, otherwise 0), and TS is by computing cosine similarity between the title and all clusters. As indicated by the results, delving into the correlation between each title and all title clusters in TS leads to better performances.

Discussion with Contrastive Learning Inspired from the self-supervised learning (He et al., 2020), the video representation learning from text is recently formulated as the contrastive learning (Li & Wang, 2020) between the visual and textual features. The features of the positive video-text pairs are pulled close while the features of negative pairs are pushed away. Following such setting, we experimented with contrastive learning on video and text via InfoNCE loss (Oord et al., 2018) on the same backbone. The top-1 accuracy achieve 72.2%, 90.1% and 62.3% on Kinetics-400, UCF101 and HMDB51 under linear evaluation protocol, respectively. The performances are lower than 72.5%, 91.8% and 64.5% of the TS in Table 2. Instead of learning the pair-wise correlation through contrastive learning, our TS mines the group-wise relationship via clustering. The results confirm that our TS is a good alternative for the video representation learning based on text information.
Table 2: Top-1/Top-5 accuracy on Kinetics-400, UCF101 and HMDB51 under linear protocol. (Training on YOVO-3M)

| Approach | Kinetics-400 |         | UCF101 |         | HMDB51 |         |
|----------|--------------|---------|---------|---------|--------|---------|
|          | Top-1       | Top-5   | Top-1   | Top-5   | Top-1  | Top-5   |
| RG       | 64.5        | 84.9    | 85.1    | 94.3    | 58.3   | 81.2    |
| TR       | 67.2        | 87.2    | 91.0    | 98.9    | 62.2   | 85.6    |
| TS-      | 71.5        | 89.1    | 91.8    | 99.0    | 64.5   | 89.7    |
| TS       | 72.5        | 89.6    |         |         |        |         |

Table 3: Performance contribution of each design in BCN

| Approach | K400 | U101 | HD51 |
|----------|------|------|------|
|          | Linear | Fine-tune | Linear | Fine-tune | Linear | Fine-tune |
| Q        | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| T        | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| QS       | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| TS       | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| QS+TS    | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| BCN_Q    | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| BCN_T    | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |
| BCN      | ✔    | ✔    | ✔    | ✔    | ✔    | ✔    |

The model is learnt on YOVO-3M and evaluated by linear and fine-tuning protocols. The “Q” and “T” denote the corresponding primary query/text supervision.

4.4 Evaluation on Bi-calibration Networks

Next, we study each design component in BCN for video representation learning. QS and TS solely exploit primary query supervision and text supervision in our framework to guide video representation learning, respectively, through single-label and multi-label classification. QS+TS performs the joint learning on the primary supervision of query and text. BCN_Q or BCN_T leverages the idea of text-to-query (T→Q) or query-to-text (Q→T) calibration to estimate t2q/q2t corrections to adjust primary query/text supervision and further boost video representation learning. BCN is our proposed Bi-Calibration framework.

Table 3 details the top-1 accuracy on Kinetics-400 (K400), UCF101 (U101) and HMDB51 (HD51) datasets under linear model and network fine-tuning protocols by considering one more factor in our BCN. Compared to QS/TS, QS+TS boosts up the top-1 accuracy under linear protocol from 71.1%/72.5% to 73.1% on K400. The similar performance trend is observed under the fine-tuning protocol. The results indicate that primary query supervision and primary text supervision are complementary to refine visual-textual connections and enhance video representation learning. T2q/q2t calibration further refine primary query/text supervision and the performance improvement of each is 1.2%/0.7% and 0.9%/0.6% under linear protocol on U101 and HD51 against QS+TS. By mutual calibration between query and text, BCN finally reaches the top-1 accuracy under linear model protocol of 74.3%, 93.6% and 67.9% on the three datasets. The corresponding performances under fine-tuning protocol are 78.7%, 97.2% and 75.4%.

Fig. 9: a 10 classes with the largest performance gain from QS to BCN under linear protocol on Kinetics-400 and the visualization of confusion matrix across the selected six categories on video representation learnt by b QS and c BCN

To verify the impact of mutual calibration via our BCN design across different categories, we further list the categories with most benefit. Figure 9a shows ten categories in Kinetics-400 which achieve the largest performance gain from QS to BCN. An interesting observation is that three pairs, i.e., kissing-hugging, eating burger-eating chips, waxing back-waxing legs, among the ten categories, are indeed fine-grained and it is challenging to distinguish one from the other in each pair. Figure 9b, c also visualizes confusion
matrix across the six categories on video representation learnt by QS and BCN, respectively. It is clear that BCN endows video representation with more discriminative power especially on the fine-grained categories.

4.5 More Analysis on Network Optimization

The selection scheme controls the switch across the calibration of t2q or q2t directions. We compute the Residual Sum of Squares (RSS) between the query/text distributions ($p^q / p^t$) and t2q/q2t corrections ($\hat{p}^q / \hat{p}^t$), and Fig. 11a depicts RSS curve with respect to the number of iterations. As expected, the RSS on both ($p^q, \hat{p}^q$) and ($p^t, \hat{p}^t$) is gradually decreased when training more iterations. Since the corrections are adopted as the supervision to optimize the probabilities, the gradients are not back-propagated to them. Thus, the results give the clue that estimated distribution of the samples with unreliable primary query or text supervision are enforced to be close to the corresponding t2q/q2t corrections, validating the impact of the two calibration procedures. Figure 11b shows the curve of top-1 accuracy on Kinetics-400 under linear protocol in two optimization stages. As illustrated by the figure, there is a performance saturation when increasing the training epochs of the first stage. Then, in the second training stage, the t2q and q2t calibrations change the training objective by adjusting the primary query/test supervision. As a result, the learnt video representation in the first training stage is no longer optimal for the new training objective. Such change of the training objective may momentarily degrade the discriminative ability of the video representation at the beginning of bi-calibration training stage, thereby resulting in the performance drop. With the optimization on new objective for several epochs, the top-1 accuracy eventually improves and reaches a higher value. The results again demonstrate the effectiveness of bi-calibration between query and text through information refinement.

Figure 10 showcases two groups of videos with the searched queries, video titles and the probability distribution on query/text ($p^q / p^t$) of the two video cases in Fig. 10 before and after the t2q/q2t calibration procedure. ocean liner in movie “Titanic” and the other describes “DJI” drone. Such query polysemy may misdirect video representation learning. With the t2q corrections in BCN, the query probabilities of the two videos are well predicted. The video about “Titanic” is highly relevant to “ocean liner” and “kissing,” and the video of “DJI” has high probability in response to “drone” and “warplane.” In contrast, the second group of videos share similar syntactic structure of title, but are related to different queries of “ice cream” and “ice lolly.” In this case, solely capitalizing on title information may affect video representation learning as well. Through query-to-text calibration, the text probabilities predicted by BCN nicely lead to different emphasis of “ice cream” or “ice lolly” and “chocolate sauce.” Fig. 12 further illustrates the probability distribution on query/text of two video cases in Fig. 10 before and after the t2q/q2t calibration procedure.
Table 4 Performance comparisons on Kinetics-400

| Approach | Pre-training | Backbone | Top-1 | Top-5 |
|----------|--------------|----------|-------|-------|
| **Supervised pre-training** | | | | |
| R(2+1)D (Tran et al., 2018) | ImageNet | Custom | 74.3 | 91.4 |
| I3D (Carreira & Zisserman, 2017) | ImageNet | Inception | 72.1 | 90.3 |
| S3D (Xie et al., 2018) | ImageNet | Inception | 74.7 | 93.4 |
| NL I3D (Wang et al., 2018b) | ImageNet | ResNet-50 | 74.9 | 91.6 |
| TSM (Lin et al., 2019) | ImageNet | ResNet-50 | 74.1 | 91.2 |
| LGD (Qiu et al., 2019) | ImageNet | ResNet-50 | 74.8 | 92.0 |
| MUSLE (Li et al., 2021a) | ImageNet | ResNet-50 | 75.1 | 92.0 |
| SlowFast (Feichtenhofer et al., 2019) | ImageNet | ResNet-50 | 75.6 | 92.1 |
| SmallBig (Li et al., 2020a) | ImageNet | ResNet-50 | 76.3 | 92.5 |
| CorrNet (Wang et al., 2020) | ImageNet | ResNet-50 | 77.2 | – |
| TDN (Wang et al., 2021a) | ImageNet | ResNet-50 | 77.5 | 93.2 |
| TAda2D (Huang et al., 2022) | ImageNet | ResNet-50 | 78.2 | 93.5 |
| VTN (Neimark et al., 2021) | ImageNet | ViT-B | 78.6 | 93.7 |
| ViTr (Zhang et al., 2021) | ImageNet | ViT-L | 80.5 | 94.6 |
| TimeSFormer (Bertasius et al., 2021) | ImageNet | TimeSFormer-L | 80.7 | 94.7 |
| MViT (Fan et al., 2021) | ImageNet | MViT-B | 81.2 | 95.1 |
| ViViT (Arnab et al., 2021) | ImageNet | ViT-L | 81.3 | 94.7 |
| MTV (Yan et al., 2022) | ImageNet | MTV-B | 81.8 | 95.0 |

**Self-supervised Pre-training (linear protocol)**

| Approach | Pre-training | Backbone | Top-1 | Top-5 |
|----------|--------------|----------|-------|-------|
| VTHCL (Yang et al., 2020a) | K400 | ResNet-50 | 37.8 | – |
| CVRL (Qian et al., 2021) | K400 | ResNet-50 | 66.1 | – |
| ρBYOL (Feichtenhofer et al., 2021) | K400 | ResNet-50 | 71.5 | – |

**Weakly-supervised Pre-training**

Linear model protocol on video representation

| Approach | Pre-training | Backbone | Top-1 | Top-5 |
|----------|--------------|----------|-------|-------|
| CPD (Li & Wang, 2020) | K400-title | ResNet-50 | 63.8 | – |
| BCN | YOVO-3M | ResNet-50 | 74.3 | 90.9 |
| BCN | YOVO-10M | ResNet-50 | 75.1 | 91.7 |

Network fine-tuning

| Approach | Pre-training | Backbone | Top-1 | Top-5 |
|----------|--------------|----------|-------|-------|
| OmniSource (Duan et al., 2020) | OmniSource | ResNet-50 | 73.6 | 91.0 |
| WS* (Ghadiyaram et al., 2019) | YOVO-3M | ResNet-50 | 75.5 | 92.0 |
| WS* (Ghadiyaram et al., 2019) | YOVO-10M | ResNet-50 | 76.1 | 92.3 |
| TR* (Stroud et al., 2020) | YOVO-3M | ResNet-50 | 75.2 | 92.1 |
| TR* (Stroud et al., 2020) | YOVO-10M | ResNet-50 | 76.2 | 92.6 |
| CPD* (Li & Wang, 2020) | YOVO-3M | ResNet-50 | 73.9 | 90.2 |
| CPD* (Li & Wang, 2020) | YOVO-10M | ResNet-50 | 75.0 | 91.3 |
| BCN | YOVO-3M | ResNet-50 | 78.7 | 94.5 |
| BCN | YOVO-10M | ResNet-50 | 79.5 | 94.8 |
| BCN | YOVO-10M | TimeSFormer-L | 81.9 | 95.3 |

“docker”. After the t2q calibration, we achieve more accurate predictions (“ocean liner” and “kissing”). In the second case, the video is first predicted into the text prototype of “pectin1” before q2t calibration. According to the data statistics, the text cluster “pectin1” reflects the cooking of pectin, while the other text cluster “pectin2” tends to describe the pectin on sale. These text clusters provide the fine-grained supervision based on different contexts of video titles. Then, the q2t calibration changes the prediction into the preciser text prototype of the query “ice cream.” Although the similar probabilities of the three title clusters of “ice cream” are presented, the title cluster “ice cream2” with the highest probability value represents the understanding of the “cooking ice cream” in the video, which is discriminated against other text prototypes of the queries.
### 4.6 Comparisons with State-of-the-Art Methods

We compare BCN with several state-of-the-art techniques on five datasets: Kinetics-400 (K400), UCF101 (U101) and HMDB51 (HD51) for scene-related action recognition, and Something-Something V1 (SS-V1) and V2 (SS-V2) for interaction-related action recognition.

Table 4 lists the top-1 and top-5 accuracy of different approaches on K400. For fair comparison, all the methods exploit RGB modality of frame for model training and "*" denotes that the models are pre-trained on our YOVO-3M/10M with official source codes. Under linear model protocol, BCN achieves comparable performances with most models built on fully-supervised ImageNet pre-training and self-supervised Kinetics pre-training. When fine-tuning the pre-trained model by BCN on K400, BCN exhibits better performances against other baselines. In particular, BCN with ResNet-50 backbone learnt on YOVO-3M obtains 78.7% top-1 accuracy, which outperforms the weakly-supervised video representation learning approach WS (Ghadiyaram et al., 2019), TR (Stroud et al., 2020) and CPD (Li & Wang, 2020) pre-trained on the same data with the same video backbone by 3.2%, 3.5% and 4.8%. Different from WS or CPD which solely capitalizes on query or title information, BCN exploits mutual calibration between the two to improve weakly-supervised video representation learning.

#### Table 5: Top-1 accuracy on UCF101 and HMDB51

| Approach                  | Pre-training | Backbone  | U101  | HD51  |
|---------------------------|---------------|-----------|-------|-------|
| **Supervised pre-training** |               |           |       |       |
| R(2+1)D (Tran et al., 2018) | K400          | Custom    | 96.8  | 74.5  |
| I3D (Carreira & Zisserman, 2017) | Img+K400      | Inception | 95.4  | 74.5  |
| S3D (Xie et al., 2018)     | Img+K400      | Inception | 96.8  | 75.9  |
| TSM (Lin et al., 2019)     | K400          | ResNet-50 | 95.9  | 73.5  |
| LGD (Qiu et al., 2019)     | Img+K400      | ResNet-50 | 96.0  | 74.7  |
| STM (Jiang et al., 2019)   | Img+K400      | ResNet-50 | 96.2  | 72.2  |
| MUSLE (Li et al., 2021a)   | K400          | ResNet-50 | 94.8  | 72.2  |
| TEA (Li et al., 2020b)     | Img+K400      | ResNet-50 | 96.9  | 73.3  |
| TDN (Wang et al., 2021a)   | Img+K400      | ResNet-50 | 97.4  | 76.3  |
| VidTr (Zhang et al., 2021) | K400          | ImageNet  | 96.7  | 74.4  |
| **Self-supervised pre-training (fine-tuning)** |               |           |       |       |
| XDC (Alwassel et al., 2020) | K400          | R(2+1)D-18 | 84.2  | 47.1  |
| SpeedNet (Benaim et al., 2020) | K400          | S3D-G    | 81.1  | 48.8  |
| CoCLR (Han et al., 2020)   | K400          | S3D-G    | 87.9  | 54.6  |
| MCN (Lin et al., 2021)     | K400          | R(2+1)D-18 | 89.2  | 58.8  |
| CVRL (Qian et al., 2021)   | K400          | ResNet-50 | 92.2  | 66.7  |
| ρBYOL (Feichtenhofer et al., 2021) | K400          | ResNet-50 | 95.5  | 73.6  |
| **Weakly-supervised Pre-training** |               |           |       |       |
| Linear model protocol on video representation |               |           |       |       |
| BCN                        | YOVO-3M       | ResNet-50 | 93.6  | 67.9  |
| BCN                        | YOVO-10M      | ResNet-50 | 95.0  | 69.8  |
| Network fine-tuning        |               |           |       |       |
| MIL-NCE (Miech et al., 2020) | HowTo100M     | ResNet-50 | 91.3  | 61.0  |
| CPD (Li & Wang, 2020)      | Instagram-300k| ResNet-50 | 92.8  | 63.8  |
| WS* (Ghadiyaram et al., 2019) | YOVO-3M       | ResNet-50 | 92.5  | 72.3  |
| WS* (Ghadiyaram et al., 2019) | YOVO-10M      | ResNet-50 | 93.6  | 73.0  |
| TR* (Stroud et al., 2020)  | YOVO-3M       | ResNet-50 | 92.3  | 73.4  |
| TR* (Stroud et al., 2020)  | YOVO-10M      | ResNet-50 | 94.1  | 74.1  |
| CPD* (Li & Wang, 2020)     | YOVO-3M       | ResNet-50 | 94.2  | 70.2  |
| CPD* (Li & Wang, 2020)     | YOVO-10M      | ResNet-50 | 95.4  | 73.5  |
| BCN                        | YOVO-3M       | ResNet-50 | 97.2  | 75.4  |
| BCN                        | YOVO-10M      | ResNet-50 | 97.6  | 76.9  |
| BCN                        | YOVO-10M      | TimeSformer-L | 98.2  | 78.7  |
Table 6  Top-1 accuracy on Something-Something V1/V2

| Approach                        | Pre-training | Backbone       | SS-V1 | SS-V2 |
|---------------------------------|--------------|----------------|-------|-------|
| **Supervised pre-training**     |              |                |       |       |
| 3D (Wang & Gupta, 2018)         | ImageNet+K400| ResNet-50      | 41.6  | –     |
| NL 3D (Wang & Gupta, 2018)      | ImageNet+K400| ResNet-50      | 44.4  | –     |
| NL 3D + gcn (Wang & Gupta, 2018)| ImageNet+K400| ResNet-50      | 46.1  | –     |
| CPNet (Liu et al., 2019)        | ImageNet     | ResNet-34      | –     | 57.7  |
| TSM (Lin et al., 2019)          | ImageNet     | ResNet-50      | 45.6  | 59.1  |
| SmallBig (Li et al., 2020a)     | ImageNet     | ResNet-50      | 48.3  | 61.6  |
| ACTION-Net (Wang et al., 2021c)| ImageNet     | ResNet-50      | –     | 62.5  |
| STM (Jiang et al., 2019)        | ImageNet     | ResNet-50      | 50.7  | 64.2  |
| TEA (Li et al., 2020b)          | ImageNet     | ResNet-50      | 51.9  | –     |
| TAda2D (Wang et al., 2022)      | ImageNet     | ResNet-50      | –     | 64.0  |
| TDN (Wang et al., 2021a)        | ImageNet     | ResNet-50      | 52.3  | 64.0  |
| MUSLE (Li et al., 2021a)        | ImageNet     | ResNet-50      | 52.5  | 65.0  |
| TimeSformer (Bertasius et al., 2021)| ImageNet    | TimeSformer-L  | –     | 62.5  |
| VidTr (Zhang et al., 2021)      | ImageNet     | ViT-L          | –     | 63.0  |
| ViViT (Arnab et al., 2021)      | ImageNet     | ViT-L          | –     | 65.4  |
| **Self-supervised pre-training**|              |                |       |       |
| BYOL (Feichtenhofer et al., 2021)| K400         | ResNet-50      | –     | 55.8  |
| MoCo (Feichtenhofer et al., 2021)| K400         | ResNet-50      | –     | 54.4  |
| **Weakly-supervised pre-training**|              |                |       |       |
| Linear model protocol on video representation |            |                |       |       |
| BCN                             | YOVO-3M      | ResNet-50      | 43.1  | 53.9  |
| BCN                             | YOVO-10M     | ResNet-50      | 44.3  | 56.2  |
| Network fine-tuning             |              |                |       |       |
| WS* (Ghadiyaram et al., 2019)   | YOVO-3M      | ResNet-50      | 44.8  | 60.1  |
| WS* (Ghadiyaram et al., 2019)   | YOVO-10M     | ResNet-50      | 47.2  | 61.8  |
| TR* (Stroud et al., 2020)       | YOVO-3M      | ResNet-50      | 44.2  | 59.8  |
| TR* (Stroud et al., 2020)       | YOVO-10M     | ResNet-50      | 46.5  | 60.9  |
| CPD* (Li & Wang, 2020)          | YOVO-3M      | ResNet-50      | 45.2  | 60.1  |
| CPD* (Li & Wang, 2020)          | YOVO-10M     | ResNet-50      | 47.1  | 61.7  |
| BCN                             | YOVO-3M      | ResNet-50      | 49.8  | 63.2  |
| BCN                             | YOVO-10M     | ResNet-50      | 51.0  | 65.0  |
| BCN                             | YOVO-10M     | TimeSformer-L  | 52.7  | 65.9  |

Such result demonstrates the advantage of our bi-calibration design. Compared to TDN (Wang et al., 2021a) and TAda2D (Huang et al., 2022), BCN learnt on YOVO-3M leads the top-1 accuracy by 1.2% and 0.5%, respectively. Executing weakly-supervised learning on YOVO-10M further improves the accuracy from 78.7% to 79.5%. We additionally implement BCN using the transformer-based video backbone TimeSformer-L (Bertasius et al., 2021) on YOVO-10M for performance comparison. BCN boosts up the top-1 accuracy on K400 from 80.7% to 81.9% under the fine-tuning protocol, which demonstrates the effectiveness of our video representation learning scheme with Transformer backbone. BCN with TimeSformer-L backbone surpasses that with ResNet-50 by 2.4% in top-1 accuracy, but requires ~21.2 times more computational cost and ~3 times more training time. Such trade-off between performance and computations will impact the choice of backbone in real-world deployments.

The performance trends on UCF101 and HMDB51 are similar with those on K400 as shown in Table 5. Under the fine-tuning evaluation protocol, BCN pre-trained on YOVO-3M surpasses CPD learnt on the same data by 3.0% and 5.2% on UCF101 and HMDB51, respectively. Compared to the model pre-trained on the Instagram-300k dataset, the performance gains (1.4% on U101 and 6.4% on HD51) contributed by our YOVO-3M on CPD are also observed. The results verify the impacts of both of our proposed datasets and BCN for weakly-supervised video representation learning.
Table 6 summarizes the performances on SS-V1 and SS-V2. With the video backbone of ResNet-50, BCN pre-trained on YOVO-10M achieves 51.0% and 65.0% top-1 accuracy on SS-V1 and SS-V2 under the fine-tuning protocol, respectively. Despite having relatively simple backbone, BCN still achieves comparable performances with the state-of-the-art supervised approaches which are deliberately designed with temporal difference module (Wang et al., 2021a) or graphical model (Li et al., 2021a) for temporal modeling. Furthermore, by training BCN with the video transformer backbone TimeSformer-L, our BCN obtains the best performances on both two datasets.

5 Conclusions

We have presented Bi-Calibration Networks (BCN), which explores the correlations between web videos and the searched queries or video titles for improving weakly-supervised video representation learning. Particularly, we study the problem from the viewpoint of refining the visual-semantic connections through mutual calibration between query and title information. To materialize our idea, we first achieve the primary query and text supervision on the query label set of one-hot labels and text label set of text prototypes, which are utilized to optimize video-to-query (v2q) and video-to-text (v2t) projections for classification. Next, the v2t/v2q projection triggers the text-to-query or query-to-text calibration, that aims to adjust primary query/text supervision to further optimize v2q/v2t projection. Extensive experiments conducted on newly-created web video datasets, i.e., YOVO-3M and YOVO-10M, validate our BCN. More remarkably, weakly-supervised pre-training BCN on YOVO-10M is superior to several techniques with fully-supervised ImageNet or Kinetics pre-training.

Data Availability

The newly-created web video datasets YOVO-3M and YOVO-10M are available at https://github.com/FuchenUSTC/BCN/tree/master/datasets. One possible issue of using the proposed two datasets is that small quantities of videos might be missing or taken down from YouTube, potentially affecting data availability in the future. The video data that support the downstream task evaluation of this research study are available in the Kinetics-400 (Carreira and Zisserman, 2017) (https://www.deepmind.com/open-source/kinetics), UCF101 (Soomro et al, 2012) (https://www.crcv.ucf.edu/data/UCF101.php), HMDB51 (Kuehne et al, 2011) (https://serre-lab.clps.brown.edu/resource), Something-Something V1 and V2 (Goyal et al, 2017) (https://developer.qualcomm.com/software/ai-datasets) datasets.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). YouTube-8M: A Large-Scale Video Classification Benchmark. arXiv preprint arXiv:1609.08675.

Agrawal, P., Carreira, J., & Malik, J. (2015). Learning to see by moving. In ICCV.

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., & Tran, D. (2020). Self-supervised learning by cross-modal audio-video clustering. In NeurIPS.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C. (2021). ViViT: A video vision transformer. In ICCV.

Avila, S., Thome, N., Cord, M., & Valle, E. (2013). de A Araujo A. The visual codeword point of view. Compute vision and image understanding: Pooling in image representation.

Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W. T., Rubinstein, M., Irani, M., & Dekel, T. (2020). SpeedNet: Learning the speediness in videos. In CVPR.

Berg, T. L., & Forsyth, D. A. (2006). Animals on web. In CVPR.

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In ICML.

Cai, Q., Wang, Y., Pan, Y., Yao, T., & Mei, T. (2020). Joint contrastive learning with infinite possibilities. In NeurIPS.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In CVPR.

Carreira, J., Noland, E., Hillier, C., & Zisserman, A. (2019). A short note on the Kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987.

Damen, D., Doughty, H., Farinella, G. M., Figler, S., Furnari, A., Kazakos, E., Moltsantis, D., Munro, J., Perrett, T., Price, W., & Wray, M. (2018). Scaling egocentric vision: The EPIC-KITCHENS Dataset. In ECCV.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In CVPR.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In ACL.

Diba, A., Sharma, V., & Gool, L. V. (2017). Deep temporal linear encoding networks. In CVPR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR.

Duan, H., Zhao, Y., Xiong, Y., Liu, W., & Lin, D. (2020). Omni-sourced Webly-supervised learning for video recognition. In ECCV.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., & Zisserman, A. (2019). Temporal cycle-consistency learning. In CVPR.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In ICCV.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In CVPR.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In ICCV.

Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., & He, K. (2021). A large-scale study on unsupervised spatiotemporal representation learning. In CVPR.

Fernando, B., Bilen, H., Gavves, E., & Gould, S. (2017). Self-supervised video representation learning with odd-one-out networks. In CVPR.

Finn, C., Goodfellow, I., & Levine, S. (2016). Unsupervised learning for physical interaction through video prediction. In NIPS.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., III HD, & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM.

Ghadiyaram, D., Feiszli, M., Tran, D., Yan, X., Wang, H., & Mahajan, D. (2019). Large-scale weakly-supervised pre-training for video action recognition. In CVPR.

Ghanem, B., Niebles, J. C., Snoek, C., Heilbron, F. C., Alwassel, H., Escorcia, V., Krishna, R., Buch, S., & Dao, C. D. (2018). The ActivityNet large-scale activity recognition challenge 2018 summary. arXiv preprint arXiv:1808.03766.
Gionis, A., Indyk, P., & Motwani, R. (1999). Similarity search in high dimensions via hashing. In *VLDB*.

Girshick, R. (2015). Fast R-CNN. In *ICCV*.

Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Frueud, I., Yaniilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., & Memisevic, R. (2017). The “something something” video database for learning and evaluating visual common sense. In *ICCV*.

Han, T., Xie, W., & Zisserman, A. (2020). Self-supervised co-training for video representation learning. In *NeurIPS*.

Har, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *CVPR*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Huang, Z., Zhang, S., Pan, L., Qing, Z., Tang, M., Liu, Z., & Jr M. H. A. (2022). TAda! Temporally-adaptive convolutions for video understanding. In *ICLR*.

Jia, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE Transactions on PAMI*, 35(1), 221–231.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*.

Jiang, B., Wang, M., Gan, W., Wu, W., & Yan, J. (2019) STM: SpatioTemporal and motion encoding for action recognition. In *ICCV*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey, *International Journal of Computer Vision*, 130, 1366–1401.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. In *ICCV*.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

Li, D., Qiu, Z., Pan, Y., Yao, T., Li, H., & Mei, T. (2021a). Representing videos as discriminative sub-graphs for action recognition. In *CVPR*.

Li, J., Zhou, P., Xiong, C., & Hoi, S. (2021b). Prototypical contrastive learning of unsupervised representations. In *ICLR*.

Li, R., Zhang, Y., Qiu, Z., Yao, T., Liu, D., & Mei, T. (2021c). Motion-focused contrastive learning of video representations. In *ICCV*.

Li, T., & Wang, L. (2020) Learning spatiotemporal features via video and text pair discrimination. arXiv preprint arXiv:2001.05691.

Li, X., Wang, Y., Zhou, Z., & Qiao, Y. (2020a). SmallBigNet: Integrating core and contextual views for video classification. In *CVPR*.

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., & Wang, L. (2020b). TEA: Temporal excitation and aggregation for action recognition. In *CVPR*.

Li, Y., Yao, T., Pan, Y., & Mei, T. (2022). Contextual transformer networks for visual recognition. *IEEE Transactions on PAMI*.

Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal shift module for efficient video understanding. In *ICCV*.

Lin, Y., Guo, X., & Lu, Y. (2021) Self-supervised video representation learning with meta-contrastive network. In *ICCV*.

Liu, X., Lee, J. Y., & Jin, H. (2019). Learning video representations from correspondence proposals. In *CVPR*.

Liu, Z., Yeh, R. A., Tang, X., Liu, Y., & Agarwala, A. (2017). Video frame synthesis using deep voxel flow. In *ICCV*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H. (2022). Video swap transformer. In *ECCV*.

Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., & Mei, T. (2019). Gaussian temporal awareness networks for action localization. In *CVPR*.

Long, F., Yao, T., Qiu, Z., Tian, X., Mei, T., & Luo, J. (2020). Coarse-to-fine localization of temporal action proposals. *IEEE Transactions on Multimedia*, 22(6), 1577–1590.

Long, F., Qiu, Z., Pan, Y., Yao, T., Luo, J., & Mei, T. (2022a). Stand-alone inter-frame attention in video models. In *CVPR*.

Long, F., Qiu, Z., Pan, Y., Yao, T., Ngo, C. W., & Mei, T. (2022b). Dynamic temporal filtering in video models. In *ECCV*.

Luo, Z., Peng, B., Huang, D. A., Alahi, A., & Fei-Fei, L. (2017) Unsupervised learning of long-term motion dynamics for videos. In *CVPR*.

Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. In *JMLR*.

Mettes, P., Thong, W., & Snoek, G. G. M. (2021). Object priors for classifying and localizing unseen actions. *International Journal of Computer Vision*, 129, 1954–1971.

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*.

Miech, A., Alayrac, J. B., Smaila, L., Laptev, I., & Sivic, J., Zisserman A. (2020). End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*.

Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*.

Mobahi, H., Collobert, R., & Weston, J. (2009). Deep learning from temporal coherence in video. In *ICML*.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Pan, Q., Gutfrid, D., Vondrick, C., & Oliva, A. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on PAMI*, 42(2), 502–508.

Neimehr, D., Bar, O., Zohar, M., & Asselmann, D. (2021). Video transformer network. In *ICCV workshop*.

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinylas, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *CVPR*.

Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. In *NIPS*.

Pan, Y., Li, Y., Luo, J., Xu, J., Yao, T., & Mei, T. (2022). Auto-captions on GIF: A large-scale video-sentence dataset for vision-language pre-training. In *ACM multimedia*.

Pathak, D., Girshick, R., Dollar, P., Darrell, T., & Hariharan, B. (2017). Learning features by watching objects move. In *CVPR*.

Qian, R., Meng, T., Gong, B., Yang, M. H., Wang, H., Belongie, S., & Cui, Y. (2021). Spatiotemporal contrastive video representation learning. In *CVPR*.

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3D residual networks. In *ICCV*.

Qiu, Z., Yao, T., Ngo, C. W., Tian, X., & Mei, T. (2019). Learning spatio-temporal representation with local and global diffusion. In *CVPR*.

Qiu, Z., Yao, T., Ngo, C. W., Zhang, X. P., Wu, D., & Mei, T. (2021). Boosting video representation learning with multi-faceted integration. In *CVPR*.

Radoslav, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *ICML*.

Saenko, K., & Darrell, T. (2008). Unsupervised learning of visual sense models for polysemous words. In *NIPS*.

Schroff, F., Cristinisi, A., & Zisserman, A. (2007). Harvesting image databases from the web. In *ICCV*.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*.
Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*.

Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *ICML*.

Stroud, J. C., Ross, D. A., Sun, C., Deng, J., Sukthankar, R., & Schmid, C. (2020). Learning video representations from textual web supervision. arXiv preprint arXiv:2007.14937.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *CVPR*.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of cluster in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 59, 411–423.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *ICCV*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *CVPR*.

Wang, H., Klauser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision, 103*(1), 60–79.

Wang, H., Tran, D., Torresani, L., & Feiszli, M. (2020). Video modeling with correlation networks. In *CVPR*.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2018). Temporal segment networks for action recognition. In *ECCV*.

Wang, L., Tong, Z., Ji, B., & Wu, G. (2021a). TDN: Temporal difference networks for efficient action recognition. In *CVPR*.

Wang, M., Xing, J., & Liu, Y. (2021b). ActionCLIP: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472.

Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *ICCV*.

Wang, X., & Gupta, A. (2018). Videos as space-time region graphs. In *ECCV*.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018b). Non-local neural networks. In *CVPR*.

Wang, X., Jabri, A., & Efros, A. A. (2019). Learning correspondence from the cycle-consistency of time. In *CVPR*.

Wang, Z., Shi, Q., & Smolic, A. (2021c). ACTION-net: Multipath excitation for action recognition. In *CVPR*.

Wei, C., Fan, H., Xie, S., Wu, C. Y., Yuille, A., & Feichtenhofer, C. (2021). Masked feature prediction for self-supervised visual pretraining. arXiv preprint arXiv:2112.09133.

Wei, D., Lim, J., Zisserman, A., & Freeman, W. T. (2018). Learning and using the arrow of time. In *CVPR*.

Wu, X., Wang, R., Hou, J., Lin, H., & Luo, J. (2021). Spatial–temporal relation reasoning for action prediction in videos. *International Journal of Computer Vision, 129*, 1484–1505.

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., & Zhuang, Y. (2019). Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*.

Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., & Schmid, C. (2022). Multiview transformers for video recognition. In *CVPR*.

Yang, C., Xu, Y., Dai, B., & Zhou, B. (2020a). Video representation learning with visual tempo consistency. arXiv preprint arXiv:2006.15489.

Yang, C., Xu, Y., Shi, J., Dai, B., & Zhou, B. (2020b). Temporal pyramid network for action recognition. In *CVPR*.

Yang, J., Feng, L., Chen, W., Yan, X., Zheng, H., Luo, P., & Zhang, W. (2020c). Webly supervised image classification with self-contained confidence. In *ECCV*.

Yao, T., Zhang, Y., Qiu, Z., Pan, Y., & Mei, T. (2021). Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*.

Yao, T., Li, Y., Pan, Y., Wang, Y., Zhang, X. P., & Mei, T. (2022a). Dual vision transformer. arXiv preprint arXiv:2207.04976.

Yao, T., Pan, Y., Li, Y., Ngo, C. W., & Mei, T. (2022b). Wave-ViT: Unifying wavelet and transformers for visual representation learning. In *ECCV*.

Zabih, R., & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV*.

Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., Chen, H., Marsic, I., & Tighe, J. (2021). VidTr: Video transformer without convolutions. In *ICCV*.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., & Lin, D. (2021). Temporal action detection with structured segment networks. *International Journal of Computer Vision, 128*, 74–95.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.