6-2021

Boosting video representation learning with multi-faceted integration

Zhaofan QIU

Yao TING

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Xiao-Ping ZHANG

Dong WU

See next page for additional authors

Author

Zhaofan QIU, Yao TING, Chong-wah NGO, Xiao-Ping ZHANG, Dong WU, and Tao MEI

Boosting Video Representation Learning with Multi-Faceted Integration

Zhaofan Qiu[†], Ting Yao[†], Chong-Wah Ngo[‡], Xiao-Ping Zhang[§], Dong Wu[†] and Tao Mei[‡]

[†] JD AI Research, Beijing, China

[‡] Singapore Management University, Singapore   [§] Ryerson University, Toronto, Canada

{zhaofanqiu, tingyao.ustc}@gmail.com, cwngo@smu.edu.sg

xzhang@ee.ryerson.ca, wudong99@gmail.com, tmei@jd.com

Abstract

Video content is multifaceted, consisting of objects, scenes, interactions or actions. The existing datasets mostly label only one of the facets for model training, resulting in the video representation that biases to only one facet depending on the training dataset. There is no study yet on how to learn a video representation from multifaceted labels, and whether multifaceted information is helpful for video representation learning. In this paper, we propose a new learning framework, MUlti-Faceted Integration (MUFI), to aggregate facets from different datasets for learning a representation that could reflect the full spectrum of video content. Technically, MUFI formulates the problem as visual-semantic embedding learning, which explicitly maps video representation into a rich semantic embedding space, and jointly optimizes video representation from two perspectives. One is to capitalize on the intra-facet supervision between each video and its own label descriptions, and the second predicts the "semantic representation" of each video from the facets of other datasets as the inter-facet supervision. Extensive experiments demonstrate that learning 3D CNN via our MUFI framework on a union of four large-scale video datasets plus two image datasets leads to superior capability of video representation. The pre-learnt 3D CNN with MUFI also shows clear improvements over other approaches on several downstream video applications. More remarkably, MUFI achieves 98.1%/80.9% on UCF101/HMDB51 for action recognition and 101.5% in terms of CIDEr-D score on MSVD for video captioning.

1. Introduction

Deep Neural Networks have been proven to be highly effective for learning vision models on large-scale datasets. To date in the literature, there are various image datasets (e.g., ImageNet [47], COCO [31], Visual Genome [25]) that include large amounts of expert labeled images for train-
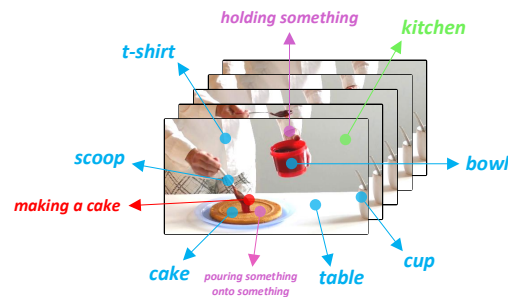

Figure 1. Example illustrating the four facets of video content: object (blue), scene (green), interaction (magenta), and action (red).

ing deep models. The well-trained models, on one hand, manifest impressive classification performances, and on the other, produce discriminative and generic representation for image understanding tasks. Compared to static 2D images, video has one more dimension (time) and is an information-intensive media with large variations and complexities. As a result, learning a powerful spatio-temporal video representation is yet a challenging problem.

Capitalizing on the high capability of deep models, one natural way to improve video representation is to acquire more video data. For example, Tran *et al.* [50] devised a widely-adopted 3D CNN, namely C3D, optimized on a large-scale Sports1M dataset and a constructed I380K dataset. Carreira *et al.* [2] built a popular pre-training dataset, i.e., Kinetics, consisting of around 300K well-annotated trimmed video clips. To expand the study in the regime of web videos which is multiple orders of magnitude larger, Ghadiyaram *et al.* [14] collected 65M web videos for pre-training 3D CNN in a weakly-supervised manner. Despite the tremendous progresses, performing learning on a specific dataset usually focuses on a particular channel of videos (e.g., action) and seldom explores other facets of videos simultaneously. Taking a video of *"making a cake"* from Kinetics dataset as an example (Figure 1), there are a wide variety of facets, ranging from object, scene, interaction, to action. Nevertheless, Kinetics, as human action dataset, mainly emphasizes the facet of action, making the learnt representation mostly aware of the action informa-

tion. A valid question is how to leverage multifaceted video information for representation learning.

This paper explores the integration of six facets, i.e., action (Kinetics [2]), event (Moments-In-Time [38]), interaction (Something-Something [15]), sport (Sports1M [14]), object (ImageNet [47]), and scene (Place365 [64]), and each facet corresponds to one dataset. We aim for a model that engages all the six facets (datasets) to learn video representation, ideally making the representation more discriminative and generic. The inherent difficulty of learning such a representation is: *how to execute effective representation learning on various datasets with different labels in a unified framework?* We propose to mitigate this issue through visual-semantic embedding learning. The basic idea is to learn a semantic space that bridges the labels from different datasets. The model that generates the space is pre-trained on a large-scale unannotated text data. The learning objective is to model the semantic relationships between labels and embed the disjoint labels into semantically-meaningful feature vectors. We project video representation into the semantic space and optimize visual-semantic embedding to enhance representation learning.

To materialize our idea, we present a new MUlti-Faceted Integration (MUFI) framework for video representation learning. Specifically, we employ off-the-shelf language models such as BERT [6] to extract textual features and build the semantic space, which is also taken as the embedding space. Each video is fed into a 3D CNN to obtain the video representation and then mapped into the embedding space. The learning of visual-semantic embedding is supervised by intra-facet embedding of a video and its class label within a dataset, and inter-facet label prediction and embedding across multiple datasets. Our MUFI framework capitalizes on the two types of supervision to jointly learn visual-semantic embedding and adjust 3D CNN through a multi-attention projection structure, and performs the whole training in an end-to-end manner. Note that the image datasets only offer the inter-facet supervision to the videos here and are not exploited as the network inputs.

The main contribution of this work is the exploration of multifaceted video content from various datasets to improve video representation learning. The novel idea leads to the elegant views of how to relate the facets of videos from different datasets, and how to consolidate various facets in a unified framework for learning, which are problems not yet fully understood. We demonstrate the effectiveness of our MUFI framework on a union of four large-scale video datasets plus two image datasets in the experiments.

2. Related Work

The early works of using Convolutional Neural Networks for video representation learning are mostly extended from 2D CNN for image classification [7, 12, 24, 43, 48,

55, 56, 65]. These approaches often treat a video as a sequence of frames or optical flow images, while overlooking the pixel-level temporal evolution across consecutive frames. To alleviate this issue, 3D CNN [21, 50] is devised to directly learn spatio-temporal representation from a short video clip via 3D convolution, which shows good transferability to several downstream tasks [27, 28, 32, 33, 34, 44]. Despite having encouraging performances, the training of 3D CNN is computationally expensive and the model size suffers from a massive growth. Later in [45, 52, 59], the 3D convolution is approximately decomposed into one 2D spatial convolution plus one 1D temporal convolution. Recently, more advanced techniques are presented for 3D CNN, including inflating 2D kernels [2], non-local pooling [57], local-and-global diffusion [46], and filter banks [36].

The aforementioned works predominantly focus on the designs of network architectures, and another direction of video representation learning is to involve more data. Ghadiyaram *et al.* [14] collect a large dataset with 65M web videos, which is multiple orders of magnitude larger than the existing video datasets. The 3D CNN in [14] is trained on such large dataset in a weakly-supervised manner and shows obvious improvements over other pre-training strategies. Later, OmniSource [8] further improves the work by utilizing weakly-labeled web images and additional web videos. Instead of treating search queries as weakly-supervised labels, the title of video in the website can also be regarded as weak supervision. In [29], the correlation between a video and its associated title is explored to learn a rich semantic video representation. Inspired by the recent advances of self-supervised learning, there are also some approaches [42, 53, 60, 61] which propose to learn video representation from unannotated video data.

In summary, our work belongs to supervised video representation learning. Different from the existing methods that focus learning on a particular facet of videos, our approach contributes by studying how to present the facets of videos from different datasets and proposing a novel video representation learning framework to consolidate various facets.

3. Multi-Faceted Integration

In this paper, we devise MUlti-Faceted Integration (MUFI) framework to integrate the rich facets from different datasets to boost video representation learning. Specifically, MUFI formulates the problem as visual-semantic embedding learning, which explicitly maps video representation into a rich semantic embedding space. We begin this section by introducing the visual-semantic embedding for video representation learning, followed by the two learning perspectives, i.e., intra-facet supervision and inter-facet supervision. Then, we propose a multi-attention projection structure to embed the visual feature. Finally, we elaborate the end-to-end optimization of our MUFI.

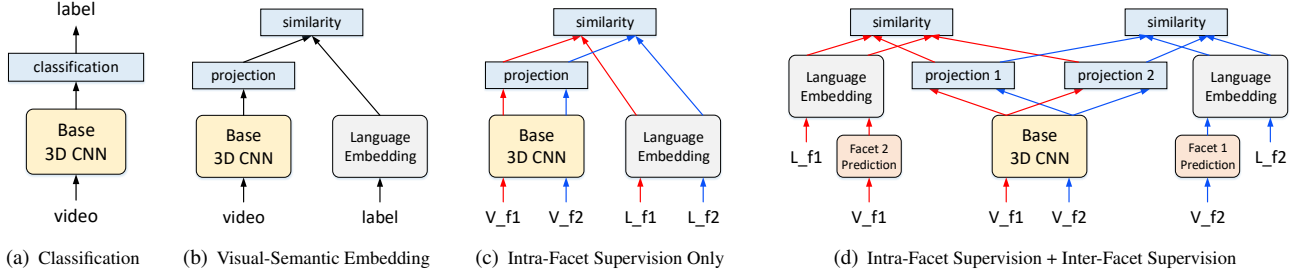| (a) Classification | (b) Visual-Semantic Embedding | (c) Intra-Facet Supervision Only | (d) Intra-Facet Supervision + Inter-Facet Supervision |

Figure 2. The schematic illustration of video representation learning by (a) classification loss; (b) visual-semantic embedding; (c) multi-faceted visual-semantic embedding with intra-facet supervision; (d) multi-faceted visual-semantic embedding with intra-facet supervision and inter-facet supervision. For the multi-faceted frameworks, the arrows of data flow from an identical facet are in the same color.

3.1. From Classification to Embedding

We firstly study the most common single-facet representation learning problem. Suppose we have a labeled video collection $\mathcal{V} = \{(v, l)\}$, where $l \in \mathcal{L}$ is the assigned label to describe the video content. The goal is to pre-train a visual encoder $\mathcal{F}(\cdot)$ on video data supervised by the semantic labels. The pre-trained video encoder can be further utilized to support several video downstream tasks. One conventional way is to optimize the encoder $\mathcal{F}(\cdot)$ with classification loss, as shown in Figure 2(a). In our case, we choose 3D CNN as visual encoder and then append a linear classifier $\mathcal{C}(\cdot)$ on top of the network to predict the label. Hence, the visual encoder and classifier are jointly optimized by Softmax Cross-Entropy (SCE) loss as

$$L_{classification}(v, l) = \text{SCE}\left(\mathcal{C}(\mathcal{F}(v)), l\right). \quad (1)$$

Instead, the multi-faceted learning involves N video collections $\mathcal{V}^n|_{n=1,...,N}$ and the corresponding label set $\mathcal{L}^n$. Eq.(1) can be straightforwardly extended by merging the label sets or establishing multiple classifiers. Nevertheless, these simple extensions treat each facet independently, but ignore the semantic relationships between labels.

Inspired by deep visual-semantic embedding [13], learning is formulated as a problem of aligning video representation and label representation in a rich semantic embedding space, as shown in Figure 2(b). Formally, the label representation is achieved by a frozen language embedding model $\mathcal{S}(\cdot)$ (e.g., BERT) pre-trained on unannotated text data. The objective of visual-semantic embedding is to minimize the distance between the projected video representation and the label representation:

$$L_{embedding}(v, l) = \text{distance}\left(\mathcal{P}(\mathcal{F}(v)), \mathcal{S}(l)\right), \quad (2)$$

where $\mathcal{P}(\cdot)$ is a projection function to embed the video representation into the semantic space. As such, the labels from different datasets share the same continuous feature space and are comparable to each other.

3.2. Intra-Facet Supervision

To tackle the challenge of multi-faceted embedding, we first introduce an intra-facet supervision that minimizes the

distance between video representation and label representation in the same facet, as illustrated in Figure 2(c). Given a video-label pair (v^n, l^n) from the nth facet, the intra-facet supervision can be formulated as minimizing the L2 distance between embedded representations:

$$L_{intra-l2}(v^n, l^n) = \|\mathcal{P}(\mathcal{F}(v^n)) - S(l^n)\|_2^2. \quad (3)$$

However, minimizing the L2 loss only tends to make the video representation close to the annotated label while ignoring other labels in the same facet. Such kind of training may affect the discriminative capability of the learnt representation in this facet, since other labels might be even closer than the annotated label. Similar phenomenon is also discussed in [13]. Here, we propose to address this issue by the recent success of contrastive learning [16, 17, 40]. The basic principle is to make positive/negative query-key pairs similar/dissimilar. In our case, the positive pair is constructed by taking the embedded visual representation $\mathcal{P}(\mathcal{F}(v^n))$ as query and the representation of the annotated label $S(l^n)$ as positive key. The representations from other labels in the same facet are regarded as the negative keys. Formally, by measuring the query-key similarity via dot product, the intra-facet loss can be formulated based on a softmax formulation:

$$L_{intra-nce}(v^n, l^n) = -\log \frac{\exp(\mathcal{P}(\mathcal{F}(v^n)) \cdot S(l^n))}{\sum_{\hat{l}^n \in \mathcal{L}^n} \exp(\mathcal{P}(\mathcal{F}(v^n)) \cdot S(\hat{l}^n))}, \quad (4)$$

which is similar as the prevailing form of contrastive loss in InfoNCE [40] but without the temperature parameter. The motivation behind Eq.(4) is to minimize the distance between a query (video representation) and its positive key (the assigned label) while remaining to be distinct to negative keys (the other labels) in the meantime.

3.3. Inter-Facet Supervision

In the intra-facet supervision, each video is only supervised by a single label from one facet (its source facet), while leaving the other information in the video unexploited. Although we can pool the multiple datasets together by leveraging the single-faceted labels of a dataset

at a time, learning is still based on single facet and multifaceted analysis is not exploited, resulting in data inefficiency. To alleviate the issue, we additionally involve the inter-facet supervision on video data to explore the potential of the other facets via transferring the pre-learnt knowledge, as illustrated in Figure 2(d). Specifically, when transferring the knowledge of the mth facet to video $v^n \in \mathcal{V}^n$, we utilize a classification model pre-trained on the mth facet to predict the category probability of the input video as $p^m(\hat{l}^m|v^n)$. Our goal is to make the embedded video representation close to the category with high probability in the mth facet. Thus, we calculate the sum of label representations weighted by their label probabilities as the pseudo "semantic representation" of video v^n as

$$S^m(v^n) = \sum_{\hat{l}^m \in \mathcal{L}^m} p^m(\hat{l}^m|v^n)S(\hat{l}^m). \quad (5)$$

The classification model can be considered as an expert "labeler" in m^{th} facet, and the prediction from this "labeler" is mapped to the semantic space. Then, the objective of inter-facet supervision is to minimize the distance between embedded video representation and pseudo representation:

$$L_{inter}(v^n) = \sum_{m \neq n} \|\mathcal{P}^m(\mathcal{F}(v^n)) - S^m(v^n)\|_2^2. \quad (6)$$

Unlike intra-facet supervision with strict labels, we do not require the embedded visual feature to become far apart from the other labels for inter-facet supervision and thus simply choose the L2 loss. By minimizing the loss in Eq.(6), the videos from any dataset can be supervised by all the facets. Moreover, our proposal also provides an elegant way to learn from image datasets by transferring the pre-trained image classifier to video data.

3.4. Multi-Attention Projection

In many cases, each facet is only related to some local regions that contain special objects or motion. Therefore, directly using one global feature (e.g., feature from global pooling layer) may lead to suboptimal results due to the noises introduced from regions that are irrelevant to the facet. In order to enable the framework to pinpoint the spatio-temporal regions that are most indicative for each facet, we devise a self-attention component plus a linear embedding layer as the projection function $\mathcal{P}^n$ for each facet. The self-attention assigns a positive weight score to each local descriptor extracted from the last convolutional layer of the 3D CNN. The score can be interpreted as the probability that the spatio-temporal region should be attended for this facet. Then the linear embedding layer maps the attended feature to the semantic space.

Technically, given the feature map $\mathcal{F}(v) \in \mathbb{R}^{T \times H \times W}$ extracted from the 3D CNN, the weight-sum of local de-
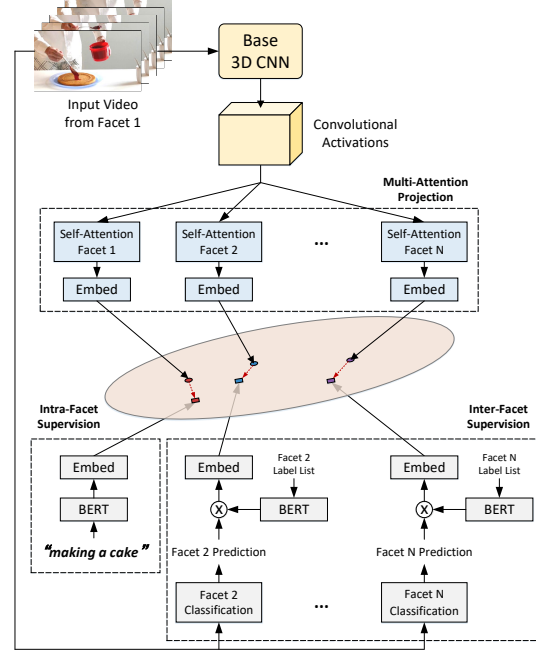

Figure 3. The overview of our MUFI framework for video representation learning. In the figure, we take the video of *"making a cake"* from facet 1 as an example. The video feature extracted from a base 3D CNN is mapped into a semantic embedding space by the multi-attention projection. The base 3D CNN and projection function is optimized in an end-to-end manner with both the intra-facet supervision and inter-facet supervision.

scriptors by self-attention can be calculated by

$$\sum_{f_{(i)} \in \mathcal{F}(v)} \varphi(w^n \cdot f_{(i)} + b^n)f_{(i)}, \quad (7)$$

where $\varphi(\cdot)$ is the softmax normalization across different spatio-temporal positions to produce the positive attention probabilities in the range of $(0, 1)$. $\{w^n, b^n\}$ are learnable parameters to distinguish the important regions for the nth facet. Please note that more complex projection functions, e.g., Multiple Layer Perception (MLP), could improve the embedding performance but meanwhile might influence the learnt representation. In view that our goal is to train the base 3D CNN for a better video representation, we choose the relatively simple multi-attention projection.

3.5. Overall Framework

Figure 3 shows the overall framework of our MUlti-Faceted Integration (MUFI), which utilizes the proposed intra-facet supervision, inter-facet supervision and multi-attention projection. Considering that the ultimate goal of MUFI is to train the base 3D CNN for video representation, we fix the pre-trained language model, embedding layer for language feature and the classification models, and only optimize the base 3D CNN and the multi-attention projection. Specifically, the language model is BERT pre-trained on

large-scale unannotated text data. The parameters of language embedding layer is achieved by Principal Component Analysis (PCA) across all the label representations. For the base 3D CNN and the multi-attention projection, we update the parameters according to the following overall loss as

$$L(v^n, l^n) = L_{intra-nce}(v^n, l^n) + L_{inter}(v^n). \quad (8)$$

Here, we empirically treat each facet in MUFI equally.

4. Experiments

4.1. Dataset and Evaluation

We train our MUFI framework on a union of six large-scale datasets: Kinetics [2], Moments-In-Time [38], Something-Something [15], Sports1M [14], ImageNet [47] and Place365 [64]. The first four are video recognition datasets focusing on action, event, interaction and sport, respectively. The last two are popular image datasets for object recognition and scene recognition.

Kinetics is one of the large-scale video benchmarks for action recognition on trimmed video clips. We utilize the latest released version of the dataset, i.e., Kinetics-700, which contains around 600K video clips from 700 action categories. **Moments-In-Time** is a large-scale video event recognition dataset with around 800K video moments (3-second short videos) from 339 events. The **Something-Something** dataset is firstly constructed in [15] to learn fine-grained human-object interaction, and then extended to Something-Something V2 recently. **Sports1M** dataset consists of about 1.13M videos annotated with 487 sports labels. Please note that some video URLs were dead when we downloaded the videos. Hence, we only employ around 1M available videos. For each raw video, we randomly sample five 5-second short clips for efficient training. **ImageNet** and **Place365** are image benchmarks with over 1M images from 1,000 object categories and 365 scene categories, respectively. These two datasets are only exploited to provide inter-facet supervision and never used as inputs to 3D CNN.

Per-facet evaluation. To thoroughly evaluate the learnt representation with respect to every facet, we construct a multifaceted evaluation metric on the classification of action (**act.**), event (**eve.**), interaction (**int.**), sport (**spo.**), object (**obj.**) and scene (**sce.**), respectively. For the first four facets, we randomly sample 50 videos of each class from the training/validation set of Kinetics-700, Moments-In-Time, Something-Something V2 and Sports1M datasets as the standalone training/validation set. The videos in these four subsets will be only exploited to evaluate the learnt representation and are not used to train MUFI. For object recognition, we utilize the **ImageNet object detection from video** (VID) [47] dataset with 3,862 video clips of 30 object categories. Please note that VID dataset provides the bounding box of each object and we only use the

most frequent object as the video label. For scene recognition, we employ **YUP++** [11] that contains 1,200 video from 20 scene categories.

Downstream task evaluation. We also verify the merit of MUFI in three downstream tasks: action recognition (UCF101 [49], HMDB51 [26] and Kinetics-400), interaction recognition (Something-Something V1 and V2) and video captioning (MSVD [3]). For fair comparison, during the training of MUFI, we remove the training videos that are duplicate of validation videos of these datasets. Two videos are judged to be duplicate if they share the same URL, or are visually very similar. More details for video deduplication will be given in the supplementary material.

4.2. Implementation Details

Language embedding. The 24-layer BERT [6] is utilized as the language embedding model pre-trained on unannotated text from BooksCorpus [67] and Wikipedia. We extract the output of the last but one layer in BERT as a sequence of 1,024-dimensional word representations. The language representation is then produced by concatenating the features max-pooled and average-pooled over all words, resulting in 2,048-dimensional feature vector. Finally, the language feature is reduced to 256 dimension by PCA to construct the embedding space.

Training strategy. Our MUFI is implemented in Caffe [22] and the weights are trained by SGD. We employ LGD-3D [46] originated from the ImageNet pre-trained ResNet-50 [19] as our backbone. During training, the dimension of input video clip is set as $16 \times 112 \times 112$, which is randomly cropped from the non-overlapped 16-frame clip resized with the short edge in $[128, 170]$. Random horizontal flipping and color jittering are utilized for augmentation.

Evaluation protocol. We exploit two evaluation protocols, i.e., linear model and network fine-tuning. In the former protocol, we directly exploit the backbone learnt by MUFI as the feature extractor, and verify the frozen representation via linear classification. Specifically, we uniformly extract 20 clip-level representation and average all the clip-level features to obtain the video-level representation. A linear SVM is trained on the training set and evaluated on each validation set. In the latter one, the pre-trained model by MUFI serves as the network initialization for further fine-tuning in downstream tasks. At inference stage, we employ the three-crop strategy in [10] that crops three 128×128 regions from each video clip.

4.3. Evaluations on MUFI w.r.t Every Facet

We first examine the effectiveness of MUFI for video representation learning with respect to every facet under linear model protocol from three perspectives: (1) video representations learnt on single/multiple facet(s) in different ways, (2) difference between MUFI and training/fine-

Table 1. Per-facet performance comparisons across video representations learnt on single/multiple facet(s) in different ways.

Method	act.	eve.	int.	spo.	obj.	sce.	avg
Video Representation Learning via Classification							
action-only (K700)	57.6	17.9	17.8	53.1	74.2	93.1	52.2
event-only (MIT)	40.4	26.0	16.3	46.1	68.6	95.3	48.7
interaction-only (SSV2)	18.9	5.6	38.6	25.6	46.6	67.1	33.7
sport-only (S1M)	34.8	11.6	10.4	67.3	66.6	88.5	46.7
4facets	51.3	22.7	37.4	62.9	74.2	95.1	57.2
4facets+multi-fc	51.4	21.0	40.8	61.1	73.5	95.2	57.1
Video Representation Learning via Visual-Semantic Embedding							
$intra_{L2}$	52.3	23.2	32.0	63.8	73.2	95.1	56.6
$intra_{nce}$	55.6	25.9	42.4	65.3	74.7	95.6	59.9
$intra_{nce}$+$inter_{video}$	56.2	26.2	44.1	65.7	74.6	94.8	60.2
$intra_{nce}$+$inter_{all}$	55.7	26.2	44.3	65.6	76.7	96.9	60.9
MUFI (+attention)	**57.8**	**27.0**	**47.8**	**67.7**	**77.8**	**97.5**	**62.6**

Table 2. Per-facet comparisons between MUFI and training/fine-tuning the network by considering one facet after another.

Method	act.	eve.	int.	spo.	obj.	sce.	avg
K700	57.6	17.9	17.8	53.1	74.2	93.1	52.2
K700+MIT	45.0	**27.3**	16.6	47.6	70.0	94.8	50.2
K700+MIT+SSV2	21.7	9.1	39.7	28.0	43.7	64.2	34.4
K700+MIT+SSV2+S1M	35.7	18.2	12.3	**68.8**	69.2	88.2	48.7
MUFI	**57.8**	27.0	**47.8**	67.7	**77.8**	**97.5**	**62.6**

tuning the network by considering one facet after another, and (3) comparisons with video representations extracted by off-the-shelf vision models.

Video representations learnt on single/multiple facet(s) in different ways. We compare the following runs for performance evaluation. The run of action-only/event-only/interaction-only/sport-only performs video representation learning via classification on Kinetics-700 (K700), Moments-In-Time (MIT), Something-Something V2 (SSV2) and Sports1M (S1M), respectively, which leverages only one facet for model training. The run, 4facets, exploits one fc layer to execute classification on the union of all annotated categories on four facets. A variant of 4facets, namely 4facets+multi-fc, separates the fc layer into 4 groups, each of which corresponds to the categories from one specific facet. Another line to learn video representation delves into visual-semantic embedding. The methods, $intra_{L2}$ and $intra_{nce}$, capitalize on intra-facet supervision of the four datasets. The former formulates the embedding learning as minimizing L2 distance of video-label pairs, while the latter exploits NCE loss in contrastive learning framework. The run of $intra_{nce}$+$inter_{video}$ and $intra_{nce}$+$inter_{all}$ further take the inter-facet supervision across four video datasets and four video datasets plus two image datasets into account for cross-view embedding learning. MUFI is the proposal in the paper that involves attention mechanism to reflect the focus of each facet.

Table 1 summarizes the result comparisons on the validation sets of different facets. Specifically, action-only/event-only/interaction-only/sport-only method achieves the highest accuracy on its corresponding facet of evaluations, but performs relatively poor on other facets. The results are expected as solely learning on one dataset biases video representation to only one facet of the training dataset. Directly treating the four datasets as a whole and implementing classification of 4facets and 4facets+multi-fc leads the accuracy by 4.9%∼23.5% on average over single facet. The results basically verify the merit of multi-faceted learning to enhance video representation. Though $intra_{L2}$ and $intra_{nce}$ in-

volve the utilization of intra-facet supervision, they are different in the way that $intra_{L2}$ exploits each video-label connection independently, and $intra_{nce}$ explores the influence across different connections. As indicated by the results, $intra_{nce}$ leads to the boost against $intra_{L2}$ by 3.3%.

The run of $intra_{nce}$+$inter_{video}$ and $intra_{nce}$+$inter_{all}$ further employ the inter-facet supervision and improve the average accuracy from 59.9% to 60.2% and 60.9%. Note that $intra_{nce}$+$inter_{video}$ explores the inter-supervision from the first four facets and emphasizes more on these facets. As a result, it is not surprising that $intra_{nce}$+$inter_{video}$ exhibits consistently better performances than $intra_{nce}$ on these facets but with slightly worse performances on the other two facets. In comparison, $intra_{nce}$+$inter_{all}$ is benefited from the inter-supervision among all the facets in both video and image datasets and outperforms $intra_{nce}$ across all facets. With the multi-attention projection, MUFI nicely balances the contribution of each facet for video representation learning and obtains a clear performance gain from 60.9% to 62.6%. More importantly, MUFI achieves the highest accuracy across all the facets, showing the potential of learning more discriminative and generalized features.

Difference between MUFI and training/fine-tuning the network by considering one facet after another. One common way in practice for video representation learning on multiple datasets/facets is to pre-train the network on a large-scale video dataset and then fine-tune the network by other datasets one by one. We take K700 as the basis to pre-train the network and fine-tune the architecture successively with MIT, SSV2 and S1M. Table 2 details the changes of accuracy as the network fine-tuning proceeds with one dataset after another. An interesting observation is that the performance gain tends to be large on the facet of dataset the most recently used for fine-tuning while the performances on other facets learnt before that drop significantly. For instance, fine-tuning the network with S1M boosts up the accuracy on the facet of sport from 28% to 68.8%. In the meantime, the accuracy on the facet of action decreases from 57.6% to 35.7%. We speculate that this may be the result of local emphasis on the latest facet and the knowledge learnt before may be partially "forgot." In contrast, MUFI explores the supervision among all the facets holistically and performs the representation learning on multiple datasets simultaneously. The results again demonstrate the advantage of our MUFI on learning generalized representation across all the facets.

Table 3. Per-facet comparisons with video representations extracted by off-the-shelf vision models.

Method	act.	eve.	int.	spo.	obj.	sce.	avg
Image-based Representation							
ResNet-50 [19]	33.1	10.5	11.1	42.5	**80.5**	89.0	44.4
SimCLR [5] (self-supervised)	31.4	10.8	12.2	42.4	75.3	89.7	43.6
MoCo v2 [18] (self-supervised)	32.0	11.2	12.1	42.7	79.4	90.0	44.5
Video-based Representation							
C3D [50]	24.6	8.6	8.2	60.1	57.6	83.9	40.5
P3D ResNet-152 [45]	35.8	11.2	11.9	60.8	74.7	90.4	47.3
R(2+1)D-34-IG65M [14]	53.1	21.1	15.8	54.5	70.5	95.2	51.7
R(2+1)D-34-IG65M+K400 [14]	54.1	20.0	14.3	50.8	67.9	92.9	50.0
SeCo [61] (self-supervised)	34.3	11.8	11.5	45.2	63.7	88.7	42.5
MUFI	**57.8**	**27.0**	**47.8**	**67.7**	77.8	**97.5**	**62.6**

Comparisons with video representations extracted by off-the-shelf vision models. Here, we include both image-based and video-based models learnt in supervised, weakly-supervised or self-supervised manner for comparison, and Table 3 lists the results. In general, ResNet-50, SimCLR and MoCo V2 all show good performances on the object facet, and MoCo V2 slightly outdoes ResNet-50 on average. Compared to C3D, P3D devises a deeper network and involves the pre-training on ImageNet, showing better performances across all the facets. R(2+1)D-34-IG65M model is trained on 65 million weakly-supervised social-media videos and hashtags, and performs fairly well with good generalization to all the facets. Further fine-tuning the model with K400, improves feature discrimination on the facet of action, but unfortunately, affects the generalization on other facets. SeCo, as a self-supervised model on video data, is superior to MoCo V2 on the motion-related facets, e.g., action, event and sport, but inferior to MoCo V2 on the appearance-related facets, e.g., object and scene. Note that because MUFI only leverages the inter-facet supervision from image data to strengthen the learning of object facet in video representation, the learning on object facet may not be utilized as fully as direct object classification in ResNet-50. As such, MUFI performs slightly worse than ResNet-50 in object facet. On other five facets, MUFI leads to significant improvements against all the off-the-shelf models.

4.4. Experimental Analysis

Visualization of attention map in each facet. Multi-attention projection is uniquely devised in MUFI to locate the most indicative spatio-temporal regions from the viewpoint of each facet. Figure 4 illustrates three video examples with the attention maps of all the six facets and the nearest label in terms of each facet in the embedding space. Overall, attention maps are expected to infer relevant video content in response to different facets. Taking the first video as an example, MUFI focuses on the regions around the Basket from action facet and pinpoints the regions of driveway with respect to the scene facet. Such multi-faceted integration equips video content understanding more comprehen-
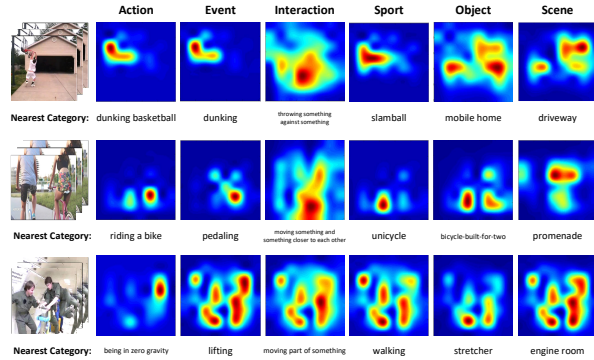

Figure 4. Video examples with per-facet attention maps and the nearest label from each facet in embedding space.
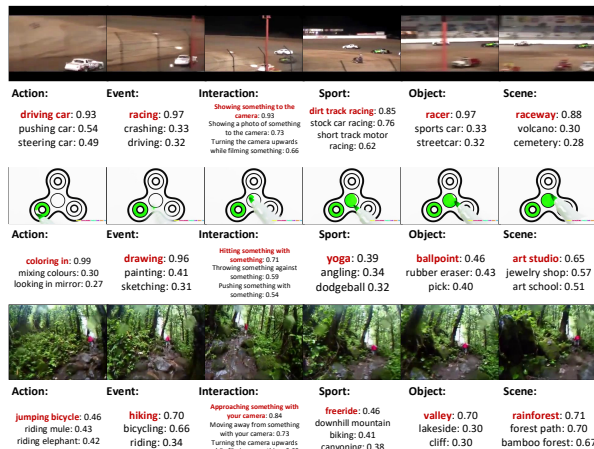

Figure 5. Video examples with the top-3 relevant labels in the embedding space from every facet.

sive and endows video representation with more power.

Examples of visual-semantic embedding. Next, we qualitatively study the learnt visual-semantic embedding by MUFI. Figure 5 showcases three video clips with the top-3 relevant labels in the embedding space from every facet and each video clip is presented by six frames. As shown in the Figure, most labels are appropriate for describing the video content from different facets. For example, driving car in action facet, racing from event standpoint, racer from object perspective and raceway with respect to scene facet are all correlated with the first video and offer more dimensions of knowledge for video representation learning.

4.5. Evaluations on Downstream Tasks

We compare with several state-of-the-art methods on three downstream tasks: action recognition on UCF101, HMDB51 and Kinetics-400, interaction recognition on Something-Something V1/V2 (SS-V1/V2), and video captioning on MSVD. Table 4 shows the comparisons on both UCF101 and HMDB51, and the performances of all the methods are reported with RGB input. MUFI+fine-tuning consistently outperforms other methods in the two datasets. In particular, MUFI+fine-tuning leads the accuracy by 1.1%

Table 4. Performance comparisons with the state-of-the-art methods with RGB input on UCF101 (3 splits) & HMDB51 (3 splits).

Method	Backbone	Pre-train	UCF101	HMDB51
Linear/Fine-tuning Protocol				
I3D [2]	BN-Inception	Kinetics-400	95.4	74.5
ARTNet [54]	BN-Inception	Kinetics-400	94.3	70.9
R(2+1)D [52]	ResNet-34	Kinetics-400	96.8	74.5
S3D-G [59]	BN-Inception	Kinetics-400	96.8	75.9
STM [23]	ResNet-50	Kinetics-400	96.2	72.2
LGD-3D [46]	ResNet-101	Kinetics-600	97.0	75.7
MUFI+linear	ResNet-50	Multi-Faceted	95.7	72.2
MUFI+fine-tuning	ResNet-50	Multi-Faceted	**98.1**	**80.9**
Zero-shot Protocol				
SAOE [37]	GoogLeNet	ImageNet	32.8	–
CPD [29]	3D ResNet50	Kinetics-210k	39.9	–
MUFI	ResNet-50	Multi-Faceted	**56.3**	**31.0**

Table 5. Performance comparisons with the state-of-the-art methods with RGB input on Kinetics-400.

Method	Backbone	Pre-train	top-1	top-5
I3D [2]	BN-Inception	ImageNet	72.1	90.3
R(2+1)D [52]	ResNet-34	Sports1M	74.3	91.4
S3D-G [59]	BN-Inception	ImageNet	74.7	93.4
NL I3D [57]	ResNet-101	ImageNet	77.7	93.3
LGD-3D [46]	ResNet-101	ImageNet	79.4	94.4
SlowFast [10]	ResNet-50	–	77.0	92.6
	ResNet-101	–	78.9	93.5
	ResNet-101+NL	–	79.8	93.9
R(2+1)D [14]	ResNet-34	IG65M	79.6	93.9
	ResNet-152	IG65M	81.3	95.1
irCSN [51]	irCSN-152	none	76.8	92.5
	irCSN-152	Sports1M	79.0	93.5
	irCSN-152	IG65M	82.6	95.3
MUFI+linear	ResNet-50	Multi-Faceted	79.0	92.3
MUFI+fine-tuning	ResNet-50	Multi-Faceted	81.1	95.1
MUFI+linear	ResNet-101	Multi-Faceted	79.8	93.2
MUFI+fine-tuning	ResNet-101	Multi-Faceted	82.3	95.3

and 5.2% over LGD-3D which uses a deeper backbone of ResNet-101. In addition to linear model and network fine-tuning protocols, we also include zero-shot protocol here, which performs classification by retrieving the nearest class based on the cosine distance between video clip and class labels in the embedding space. MUFI is again superior to SAOE and CPD. The results basically verify the generalization of the embedding space learnt by MUFI. Table 5 summarizes the performances on Kinetics-400 dataset. Similarly, the performance gain of MUFI+fine-tuning against LGD-3D is 2.9%, based on the same backbone of ResNet-101. MUFI+fine-tuning exhibits better performances than R(2+1)D-ResNet-152 with deeper network training on a much larger dataset of IG65M and performs comparable to irCSN trained on IG65M.

Table 6 and Table 7 list the comparisons on SS-V1/V2 and MSVD for interaction recognition and video captioning, respectively. The top-1 accuracy of MUFI+fine-tuning achieves 51.2% and 64.8% on SS-V1 and SS-V2, making the absolute improvement over the best competitor STM by 0.5% and 0.6%. For video captioning, we exploit MUFI model to extract video representation and feed the representation into a transformer-based encoder-decoder struc-

Table 6. Performance comparisons with state-of-the-art methods with RGB input on Something-Something V1/V2 (SS-V1/V2).

Method	Pre-train	SS-V1		SS-V2	
		Top-1	Top-5	Top-1	Top-5
NL I3D+GCN [58]	ImageNet+K400	46.1	76.8	–	–
TSM [30]	ImageNet+K400	47.2	77.1	63.4	88.5
bLVNet-TAM [9]	ImageNet	48.4	78.8	61.7	88.1
ABM-C-in [66]	ImageNet	49.8	–	61.2	–
I3D+RSTG [39]	ImageNet+K400	49.2	78.8	–	–
GST [35]	ImageNet	48.6	77.9	62.6	87.9
STDFB [36]	ImageNet	50.1	79.5	–	–
STM [23]	ImageNet	50.7	80.4	64.2	89.8
MUFI+linear	Multi-Faceted	45.1	73.4	55.2	81.8
MUFI+fine-tuning	Multi-Faceted	**51.2**	**81.0**	**64.8**	**90.0**

Table 7. Performance comparisons with state-of-the-art methods for video captioning on MSVD. (Evaluation metrics: BLEU@4 (**B@4**), METEOR (**M**), ROUGE-L (**R**) and CIDEr-D (**C**)).

Method	Backbone	B@4	M	R	C
MARN [41]	ResNet-101+C3D	48.6	35.1	71.9	92.2
OA-BTG [62]	ResNet-200+Mask-RCNN	**56.9**	36.2	-	90.6
GRU-EVE [1]	IncResV2+C3D+YOLO	47.9	35.0	71.5	78.1
MGSA [4]	IncResV2+C3D	53.4	35.0	-	86.7
POS+VCT [20]	IncResV2+C3D	52.5	34.1	71.3	88.7
ORG-TRL [63]	IncResV2+C3D+FasterRCNN	54.3	36.4	73.9	95.2
MUFI	ResNet-50	**56.9**	**37.4**	**74.2**	**101.5**

ture to produce the sentence. As shown in the table, video representation learnt by MUFI leads to a CIDEr-D score boost of 6.3% over the state-of-the-art model of ORG-TRL which fuses multiple video features extracted by Inception-ResNet-v2, C3D and Faster RCNN. This confirms the advantage of integrating multi-faceted information into video representation learning in our MUFI.

5. Conclusion

We have presented MUlti-Faceted Integration (MUFI) framework, which explores various facets in videos to boost representation learning. Particularly, we study the problem from the viewpoint of integrating the knowledge from different datasets to represent multi-faceted information of a video. We first exploit the pre-trained language model to extract textual features of labels and build the semantic space across all the datasets as the embedding space. Each video representation is then mapped into the embedding space. Both intra-facet connection between a video and its own labels, and inter-facet predictions from other datasets are taken into account as multi-faceted supervision to optimize the visual-semantic embedding and eventually improve video representation. Experiments conducted on a union of four large-scale video datasets and two image datasets validate our proposal. More remarkably, our MUFI constantly obtains superior results over state-of-the-art methods in several downstream tasks.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*, 2019. 8

[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 2, 5, 8

[3] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 5

[4] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *AAAI*, 2019. 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 7

[6] 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. 2, 5

[7] Ali Diba, Vivek Sharma, and Luc Van Gool. Deep temporal linear encoding networks. In *CVPR*, 2017. 2

[8] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, 2020. 2

[9] Quanfu Fan, Chun-Fu Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In *NeurIPS*, 2019. 8

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 5, 8

[11] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Temporal residual networks for dynamic scene recognition. In *CVPR*, 2017. 5

[12] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 3

[14] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019. 1, 2, 5, 7, 8

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2, 5

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 3

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 7

[20] Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. Joint syntax representation learning and visual cue translation for video captioning. In *ICCV*, 2019. 8

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. on PAMI*, 35(1):221–231, 2013. 2

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 5

[23] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *ICCV*, 2019. 8

[24] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 1

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 5

[27] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *ECCV*, 2018. 2

[28] Dong Li, Ting Yao, Zhaofan Qiu, Houqiang Li, and Tao Mei. Long short-term relation networks for video action detection. In *ACM MM*, 2019. 2

[29] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*, 2020. 2, 8

[30] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 8

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

[32] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, 2019. 2

[33] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Learning to localize actions from moments. In *ECCV*, 2020. 2

[34] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Tao Mei, and Jiebo Luo. Coarse-to-fine localization of tempo-

ral action proposals. *IEEE Trans. on MM*, 22(6):1577–1590, 2019. 2

[35] Chenxu Luo and Alan L. Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *ICCV*, 2019. 8

[36] Brais Martínez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe. Action recognition with spatial-temporal discriminative filter banks. In *ICCV*, 2019. 2, 8

[37] Pascal Mettes and Cees GM Snoek. Spatial-aware object embeddings for zero-shot localization and classification of actions. In *ICCV*, 2017. 8

[38] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Trans. on PAMI*, 42(2):502–508, 2019. 2, 5

[39] Andrei Liviu Nicolicioiu, Iulia Duta, and Marius Leordeanu. Recurrent space-time graph neural networks. In *NeurIPS*, 2019. 8

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[41] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *CVPR*, 2019. 8

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

[43] Zhaofan Qiu, Ting Yao, and Tao Mei. Deep quantization: Encoding convolutional activations with deep generative model. In *CVPR*, 2017. 2

[44] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Trans. on MM*, 20(4):939–949, 2017. 2

[45] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017. 2, 7

[46] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *CVPR*, 2019. 2, 5, 8

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1, 2, 5

[48] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 2

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. *CRCV-TR-12-01*, 2012. 5

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 2, 7

[51] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 8

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2, 8

[53] Jinpeng Wang, Yiqi Lin, Andy J Ma, and Pong C Yuen. Self-supervised temporal discriminative learning for video representation learning. *arXiv preprint arXiv:2008.02129*, 2020. 2

[54] Limin Wang, Weiqi Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, 2018. 8

[55] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[56] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. on PAMI*, 2018. 2

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 8

[58] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 8

[59] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 2, 8

[60] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*, 2020. 2

[61] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*, 2021. 2, 7

[62] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *CVPR*, 2019. 8

[63] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020. 8

[64] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on PAMI*, 40(6):1452–1464, 2017. 2, 5

[65] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. A key volume mining deep framework for action recognition. In *CVPR*, 2016. 2

[66] Xinqi Zhu, Chang Xu, Langwen Hui, Cewu Lu, and Dacheng Tao. Approximated bilinear modules for temporal modeling. In *ICCV*, 2019. 8

[67] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 5