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-
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 training in an end-to-end manner. Note that the image datasets only offer the inter-facet supervision to the videos training in an end-to-end manner. 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 predominately 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.
3.1. From Classification to Embedding

We firstly study the most common single-facet representation learning problem. Suppose we have a labeled video collection \( V = \{ (v, l) \} \), where \( l \in L \) is the assigned label to describe the video content. The goal is to pre-train a visual encoder \( F(\cdot) \) 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 \( F(\cdot) \) 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 \( C(\cdot) \) 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

\[
L_{\text{classification}}(v, l) = \text{SCE}(C(F(v)), l). \quad (1)
\]

Instead, the multi-faceted learning involves \( N \) video collections \( V^n | n = 1, \ldots, N \) and the corresponding label set \( L^n \). 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 \( S(\cdot) \) (e.g., BERT) pre-trained on unannotated text data. The objective of visual-semantic prediction is to minimize the distance between the projected video representation and the label representation:

\[
L_{\text{embedding}}(v, l) = \text{distance}(P(F(v)), S(l)), \quad (2)
\]

where \( P(\cdot) \) 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 \( n^{th} \) facet, the intra-facet supervision can be formulated as minimizing the L2 distance between embedded representations:

\[
L_{\text{intra}}(v^n, l^n) = \|P(F(v^n)) - S(l^n)\|^2_{L^2}. \quad (3)
\]

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 \( P(F(v^n)) \) 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:

\[
L_{\text{intra-nce}}(v^n, l^n) = -\log \frac{\exp(P(F(v^n)) \cdot S(l^n))}{\sum_{l^{\prime \prime} \in L^n} \exp(P(F(v^n)) \cdot S(l^{\prime \prime}))}, \quad (4)
\]

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 $m^{th}$ facet to video $v^n \in \mathcal{V}^n$, we utilize a classification model pre-trained on the $m^{th}$ 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 $m^{th}$ facet. Thus, we calculate the sum of label representations weighted by their label probabilities as the pseudo “semantic representation” of video $v^n$ as

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

(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:

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

(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 $P^n$ 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 $F(v) \in \mathbb{R}^{T \times H \times W}$ extracted from the 3D CNN, the weight-sum of local descriptors is

$$\phi(w^n \cdot f(i) + b^n)f(i),$$

(7)

where $\phi(\cdot)$ 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 $n^{th}$ 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 (MUF1), which utilizes the proposed intra-facet supervision, inter-facet supervision and multi-attention projection. Considering that the ultimate goal of MUF1 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

$$L(v^n, l^n) = L_{intr-nce}(v^n, l^n) + L_{inter}(v^n).$$

(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 [62]. 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 × 112 × 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-
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, intraL2 and intraece, 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 intraece+intervideo and intraece+interall 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 the utilization of intra-facet supervision, they are different in the way that intraL2 exploits each video-label connection independently, and intraece explores the influence across different connections. As indicated by the results, intraece leads to the boost against intraL2 by 3.3%.

The run of intraece+intervideo and intraece+interall further employ the inter-facet supervision and improve the average accuracy from 59.9% to 60.2% and 60.9%. Note that intraece+intervideo explores the inter-supervision from the first four facets and emphasizes more on these facets. As a result, it is not surprising that intraece+intervideo exhibits consistently better performances than intraece on these facets but with slightly worse performances on the other two facets. In comparison, intraece+interall is benefited from the inter-supervision among all the facets in both video and image datasets and outperforms intraece 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/finetuning 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 1. Per-facet performance comparisons across video representations learnt on single/multiple facet(s) in different ways.

| Method | act. | eve. | int. | spo. | obj. | sec. | 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 |

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

| Method | act. | eve. | int. | spo. | obj. | sec. | 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 |
Table 3. Per-facet comparisons with video representations extracted by off-the-shelf vision models.

| Method     | act. | eve. | int. | spo. | obj. | sce. | avg  |
|------------|------|------|------|------|------|------|------|
| 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

| Method     | act. | eve. | int. | spo. | obj. | sce. | avg  |
|------------|------|------|------|------|------|------|------|
| 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 |
| MUIF       | 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 devides 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 MUIFI 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, MUIFI performs slightly worse than ResNet-50 in object facet. On other five facets, MUIFI 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 MUIFI 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, MUIFI 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 comprehensive and endows video representation with more power.

Examples of visual-semantic embedding. Next, we qualitatively study the learnt visual-semantic embedding by MUIFI. 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. MUIFI+fine-tuning consistently outperforms other methods in the two datasets. In particular, MUIFI+fine-tuning leads the accuracy by 1.1%
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 structure 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.

**Acknowledgments.** This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600.
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 Guttridge, 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. Spatial-aware object representations. In *CVPR*, 2018. 2

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Pyramidal recurrent convolutional networks. In *ICCV*, 2019. 2, 7

[41] Xinqi Zhu, Chang Xu, Langwen Hui, Cewu Lu, and Dacheng Tao. Action recognition with spatial-temporal discriminative filter banks. In *ICCV*, 2019. 2, 5

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

[43] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal aggregation for a real-world video dataset. *CVPR*, 2019. 8

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

[45] 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

[46] 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

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

[48] 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

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

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

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

[52] 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

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

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

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

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

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

[58] 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

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

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

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

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

[63] 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

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

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

[66] 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