Learning from Semantic Alignment between Unpaired Multiviews for Egocentric Video Recognition

Qitong Wang[1], Long Zhao[2], Liangzhe Yuan[2], Ting Liu[2], Xi Peng[1]

[1]University of Delaware, [2]Google Research

{wqtwjt,xipeng}@udel.edu   {longzh,lzyuan,liuti}@google.com

Abstract

We are concerned with a challenging scenario in unpaired multiview video learning. In this case, the model aims to learn comprehensive multiview representations while the cross-view semantic information exhibits variations. We propose Semantics-based Unpaired Multiview Learning (SUM-L) to tackle this unpaired multiview learning problem. The key idea is to build cross-view pseudo-pairs and do view-invariant alignment by leveraging the semantic information of videos. To facilitate the data efficiency of multiview learning, we further perform video-text alignment for first-person and third-person videos, to fully leverage the semantic knowledge to improve video representations. Extensive experiments on multiple benchmark datasets verify the effectiveness of our framework. Our method also outperforms multiple existing view-alignment methods, under the more challenging scenario than typical paired or unpaired multimodal or multiview learning. Our code is available at https://github.com/wqtwjt1996/SUM-L.

1. Introduction

Recent years have witnessed a tremendous increase in the research regarding egocentric video understanding and learning [22, 26, 20, 41, 29, 30, 14, 42, 7, 31, 46, 9]. In parallel, a recent study [22] verifies that the activity of third-person videos in fact positively inform first-person video learning. Early works [39, 51] of egocentric video learning and understanding pay a lot of attention to paired multiview data representation learning. However, synchronized first-person and third-person video pairs are difficult to collect. This severely limits the data scale and scope of multiview learning for egocentric data.

Recently, egocentric video datasets [4, 5, 10] with larger scale have become available. However, since head-mounted cameras were not available until recent years, very few first-person videos have paired third-person videos [15].
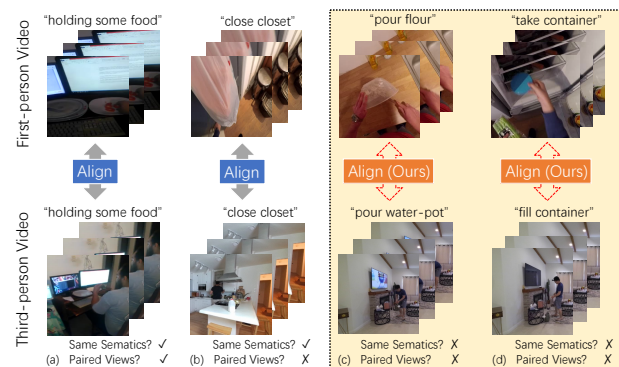

Figure 1. (a) The typical multiview video learning method learns view-invariant representations from paired first-person and third-person videos. (b) Recent multimodal study [19] is interested in the alignment between unpaired samples which share identical semantics. In this paper, we study a more challenging scenario than existing works: unpaired multiview samples alignment with partially similar semantics, such as videos with (c) the same verb only, or (d) the same object only. The verb-object phrases above the video clips denote the semantic information of videos.

Fortunately, there exist third-person videos from different datasets [12, 16]. It is worth noting that unpaired third-person videos are relatively easier to capture and obtain compared with paired ones. Therefore, in this paper, we study an interesting yet seldom investigated problem: how to leverage unpaired third-person videos to help egocentric video learning?

Unpaired multiview learning is rarely studied. While recent works [28, 27, 38, 53] study the scenario where paired modalities are missing. But they do not leverage the potential of unpaired data from diverse datasets. Moreover, to align the unpaired multimodal samples, Kundu *et al.* [19] propose one relation distillation method to align the unpaired samples. They assume that unpaired data need to share identical semantic information. However, in the unpaired first-person and third-person video learning scenario, for the vast majority of egocentric videos, it is almost impossible to mine the third-person samples with identical

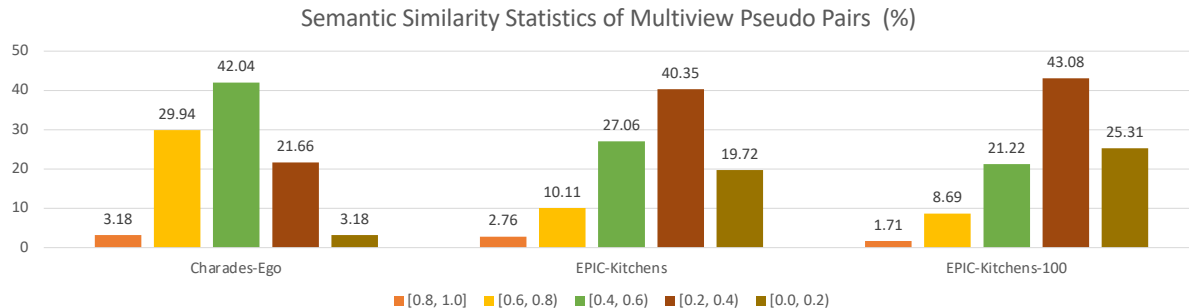Semantic Similarity Statistics of Multiview Pseudo Pairs (%)

Figure 2. Statistics of semantic similarities between the first-person samples in [39, 4, 5] and the most similar single-agent third-person samples they can find from the LEMMA [15] dataset. The semantic similarity is the cosine distance between textual phrase vectors of videos encoded by a large language model [34]. Each bar in the figure represents the percentage of action classes that fall within one semantic similarity range (*e.g*., [0.8, 1.0]) in an egocentric dataset compared to all action classes in this dataset. In all of the three popular egocentric datasets, very few first-person videos are able to find third-person videos with identical semantics. For reference, in the studies for paired multiview data [43, 39, 51, 54], all the semantic similarities of cross-view data pairs are (or very close to) "1.0".

semantics. The only alternatives are *partially semantics-similar* unpaired third-person videos. For example, for one egocentric video clip with "cut watermelon" semantics, the most similar third-person video is those with "cut lemon" (same verb) or "eat watermelon" (same object) semantics. Examples and comparisons between our scenario and typical paired or unpaired multiview or multimodal learning are shown in Figure 1, and we display comprehensive semantics similarity statistics of multiview data studied in this paper (see Figure 2). Therefore, how to employ the unpaired and partially semantics-similar third-person views to help first-person view learning is an open problem. Intuitively, compared with previous research, our setting is more challenging since most of the first-person videos only share partially similar semantics with third-person views.

In this paper, we are interested in view-invariant alignment in the partially semantics-similar pseudo-pair setting. We propose Semantics-based Unpaired Multiview Learning (SUM-L), where multiview pseudo-pairs with high similarities are aligned in a semantics-aware manner. In our method, we also employ video-text alignment to permit all first-person videos to obtain knowledge from samples with different views or modalities. We highlight that our SUM-L employs the large language model [34] to help align cross-view and cross-modality data. This is different from [22], which proposes to distill beneficial knowledge from unpaired third-person videos [16], mostly with the help of the off-the-shelf image recognition model [6] and hand-object detector [37]. Moreover, our method is better than existing view-invariant alignment methods, such as typical contrastive learning [44, 43, 50, 54] and triplet loss based learning [35, 51] in the new setting, since they naively reduce multiview feature distance, which leads to learning suboptimal first-person representations.

To summarize, our contribution is three-fold:

- We study a new problem: in the partially semantics-similar multiview setting, how to leverage unpaired third-person videos to help first-person video learning.

- We propose SUM-L for this new problem, which is more effective than existing view-alignment methods in our unpaired multiview learning scenario.

- Experiments in standard benchmark datasets including Charades-Ego [39], EPIC-Kitchens [4], and EPIC-Kitchens-100 [5] validate the competitive performance of our methods over existing view-alignment methods including typical contrastive learning and triplet loss based learning.

2. Related Work

2.1. Egocentric Video Learning

The study of egocentric video learning has attracted a large volume of attention in recent years. To verify the quality of model learning, there are multiple egocentric-related tasks, including action recognition [29, 42, 46], video summarization [26, 20], engagement detection [41], scene affordance [30], gaze prediction [14], video question answering [7], and activity anticipation [31, 9]. Nevertheless, aforementioned studies utilize knowledge from first-person videos only [26, 20, 41, 14, 42, 31, 9], or make use of extra supervision from off-the-shelf models [29, 30], or naively pretrain on large-scale datasets [7, 46]. More specifically, these methods do not consider tackling the domain mismatch between different multiview datasets or abstaining from extra supervision. Recently, Li *et al*. [22] try to discover latent signals in the third-person videos which

are predictive of key egocentric-specific properties. Similar to [29, 30], however, this method heavily relies on the knowledge of off-the-shelf image or video models.

2.2. First-person Video Representation Learning with Joint Third-person Videos

The paired first-person and third-person videos, as descriptions of the same actions from different perspectives, are potential to provide beneficial knowledge to each other, yielding more generalizable joint representations. Recent studies [39, 40] propose paired first-person and third-person datasets and learn joint representations of multiview videos. They show the potential of transferring knowledge from third-person videos to first-person videos. In order to extract more useful joint knowledge, Yu *et al*. [52] make use of the shared attention regions between different views with spatial constraints, in a self-supervised and simultaneous manner. However, nowadays very few first-person videos have paired third-person videos [15]. Moreover, we find that existing multiview methods [44, 43, 35, 51] learn suboptimal first-person representations from unpaired third-person videos, that are relatively easier to capture and obtain compared with paired data. Instead, our methodology can better distill unpaired third-person video knowledge to first-person videos, and it is more efficient and applicable than existing view-alignment methods [44, 43, 35, 51].

2.3. Unpaired Multimodal Learning

Due to the relatively difficult acquisition of paired multimodal samples and the limited number of them, unpaired multimodal learning recently has attracted the interest of some researchers. Kundu *et al*. [19] treat unpaired 3D pose learning as a self-supervised adaptation problem that aims to transfer the task knowledge from a labeled paired source domain to a completely unpaired target domain. Next, it proposes the relation distillation method to align the unpaired target videos and unpaired 3D pose sequences. However, this work makes the assumption that there exist semantically identical cross-modal data in the unpaired target domain, which is almost unavailable in unpaired first- and third-person video representation learning. Therefore, in this paper, we study a more challenging problem: how to make unpaired multiview alignment in an incompletely semantics-similar situation.

3. Methodology

In the following sections, we first describe our pseudo-pair construction method (Sec. 3.1). Then we introduce the traditional contrastive framework for paired multiview or multimodal data (Sec. 3.2). Next, we introduce our proposed cross-view and cross-modal methods (Sec. 3.3). Finally, we present our full training and evaluation pipeline in Sec. 3.4.

3.1. Semantics-based Pseudo-pair Mining

Different from the previous joint multiview learning approaches [39, 40, 51], first-person and third-person videos are not automatically paired in the cross-dataset setting. Hence, before model training, it is essential to construct pseudo multiview pairs.

Every first-person and third-person video has one unique textual narration, which describes the key semantic information of it. Therefore, we employ the large language model [34] to encode textual narrations of multiview videos to vectors. Then for each first-person video, we mine the third-person video with the highest textual semantic similarity from the third-person video pool to construct the pseudo-pair. We compute the cosine distance between cross-view video textual narrations as the criterion of semantic similarity between first-person and third-person video clips:

$$sim_{(\mathbf{z_f}, \mathbf{z_t})} = \frac{\mathbf{d_f} \cdot \mathbf{d_t}}{\|\mathbf{d_f}\| \cdot \|\mathbf{d_t}\|}, \tag{1}$$

where $\mathbf{d_f}$ and $\mathbf{d_t}$ denote textual narration vectors of first-person and third-person videos, respectively.

3.2. Typical Multiview Alignment

Existing contrastive-based multiview [43, 25, 54] or multimodal [21, 49, 32, 48] methods propose to learn the view-invariant representations via the alignment of the features. Specifically, they learn to align features between different views or modalities by minimizing the distance between paired samples and maximizing the distance between negative (unpaired) samples.

Given a batch of N normalized first-person features $\{\mathbf{z_f^1}, ..., \mathbf{z_f^N}\}$ and third-person features $\{\mathbf{z_t^1}, ..., \mathbf{z_t^N}\}$, these methods calculate the contrastive multiview Info-NCE [43, 33] loss as:

$$L_c = L_c^{(f,t)} + L_c^{(t,f)}, \tag{2}$$

$$L_c^{(v_1,v_2)} = -\frac{1}{N} \sum_{i=1}^{N} log \left[\frac{exp\left(\langle \mathbf{z_{v_1}^i}, \mathbf{z_{v_2}^i} \rangle / \tau \right)}{exp\left(\langle \mathbf{z_{v_1}^i}, \mathbf{z_{v_2}^i} \rangle / \tau \right) + S_i} \right], \tag{3}$$

$$S_i = \sum_{j \in [1,N], j \neq i} exp\left(\langle \mathbf{z_{v_1}^i}, \mathbf{z_{v_2}^j} \rangle / \tau \right), \tag{4}$$

where $\mathbf{v_1}$ and $\mathbf{v_2}$ are two views and τ is the temperature.

For an input video clip $\mathbf{x_{v_k}}$, its normalized feature representation is extracted by the encoder $f_{v_k}(\cdot)$ following the projection network $h_{v_k}(\cdot)$ as:

$$\mathbf{z_{v_k}} = \frac{h_{v_k}\left(AvgPool\left(f_{v_k}\left(\mathbf{x_{v_k}}\right)\right)\right)}{\|h_{v_k}\left(AvgPool\left(f_{v_k}\left(\mathbf{x_{v_k}}\right)\right)\right)\|_2}. \tag{5}$$

The optimization of contrastive learning methods is often computationally expensive, which requires large batch
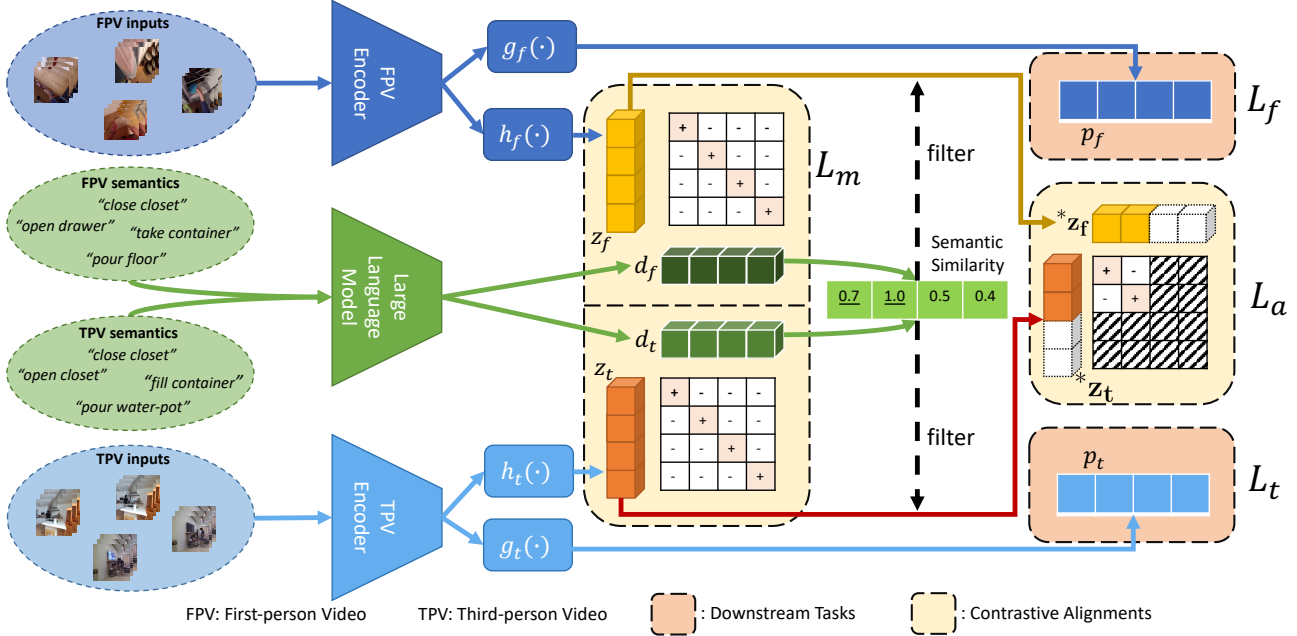
Figure 3. Illustration of our framework. First, one batch (batch size = 4) of multiview pseudo-pairs is built from unpaired first-person and third-person videos. The pseudo-pairs are built based on mining the most semantics-similar third-person video for every first-person video. During training, the global features for multiview alignment (z_f, z_t) are extracted by their corresponding encoders following the projection networks (h_f, h_t). In addition, textual features (d_f, d_t) are extracted by a large language model [34] from textual narrations of first-person and third-person videos. The semantic similarity between d_f and d_t is calculated to filter out the multiview pseudo-pairs with low semantic similarity. Then the multiview pairs with high semantic similarity are employed to learn the view-invariant representations with the contrastive learning method. To further improve data efficiency, we employ all the first-person and third-person videos in the batch to learn contrastive multimodal relations. Finally, the task-specific heads (g_f, g_t) for both first-person and third-person videos are trained to make predictions for their corresponding downstream tasks. During testing, we only use the first-person encoder and task-specific head (g_f) to make first-person video predictions.

sizes and a large volume of training epochs. To tackle this issue, Yeh *et al*. [50] refine the InfoNCE loss by removing the positive term from the denominator in Equation 3 and significantly improve the learning efficiency. To be specific, it rewrites Equation 3 as:

$$L_{dc}^{(v_1,v_2)} = -\frac{1}{N} \sum_{i=1}^{N} log \left[exp \left(\langle \mathbf{z}_{\mathbf{v_1}}^{\mathbf{i}}, \mathbf{z}_{\mathbf{v_2}}^{\mathbf{i}} \rangle / \tau \right) / S_i \right].$$

(6)

In the context of our unpaired multiview learning, we argue that directly employing the above vanilla contrastive learning method is an obstacle that hinders further acquiring knowledge from third-person videos. On the one hand, in the unpaired first-person and third-person video multiview representation learning, very few first-person views have the third-person views which share the same semantics. On the other hand, typical contrastive multiview representation learning is designed to minimize the feature distance between samples with identical semantics, *e.g.*, recordings from camera views for the same objects [18], images with different visuals belonging to the same class [17], one image applying multiple independent [2] or asymmetric [11]

data augmentations, *etc*. Naively reducing multiview feature distances in our unpaired multiview learning will lead to learning suboptimal first-person video representations.

3.3. Semantics-based Multiview Learning

To address the above issue, we propose a representation learning method, namely Semantics-based Unpaired Multiview Learning. We employ contrastive alignment as a tool to align the unpaired multiview videos. Different from existing view-alignment methods, our proposed method introduces the following two novel alignments: (1) Semantics-based Cross-view Alignment, which employs the semantic information from textual narrations of multiview videos to align high-similar unpaired data; (2) Video-text Modal Alignment, which permits all first-person video obtaining knowledge from samples with different views or modalities. Next, we will introduce our framework step-by-step.

Semantics-based Cross-view Alignment. We believe that pseudo-pairs with high semantic similarities can better help learn view-invariant information in first-person videos. Therefore, given cross-view pseudo-pairs $(\mathbf{z_f}, \mathbf{z_t})$ in one

batch, we pick pseudo-pairs $({}^*\mathbf{z_f}, {}^*\mathbf{z_t})$ which have relatively higher textual semantic similarities. Then the first-person and third-person video alignment loss (L_a) is defined as the the decoupled contrastive learning loss [50] of high semantic similarity pseudo-pairs:

$$L_a = L_{dc}^{({}^*\text{f}, {}^*\text{t})} + L_{dc}^{({}^*\text{t}, {}^*\text{f})}. \qquad (7)$$

Besides, we hypothesize that the employed large language model [34] is a strong teacher: by encoding the textual descriptions of videos into vectors, the similarity between vectors implicitly provides a reference to measure the semantic similarity between videos. To make better use of these textual descriptions of first-person and third-person videos, inspired by [50], we further introduce a semantic-based weighting function for the positive pseudo-pairs. More specifically, we first rewrite our cross-view video alignment loss L_a in Equation 7 as L_{aw}, where:

$$L_{aw} = L_{aw}^{({}^*\text{f}, {}^*\text{t})} + L_{aw}^{({}^*\text{t}, {}^*\text{f})}. \qquad (8)$$

Then given views $({}^*\text{f}, {}^*\text{t})$, we have:

$$L_{aw}^{({}^*\text{f}, {}^*\text{t})} = \frac{1}{N} \sum_{i=1}^{N} \left[log\left(S_i\right) - \omega\left({}^*\mathbf{d_f^i}, {}^*\mathbf{d_t^i}\right) \cdot \left\langle {}^*\mathbf{z_f^i}, {}^*\mathbf{z_t^i} \right\rangle / \tau \right], \qquad (9)$$

where the definition of S_i is the same as Equation 4. Here we implement the weighting function as:

$$\omega\left({}^*\mathbf{d_f^i}, {}^*\mathbf{d_t^i}\right) = \frac{exp\left(\left\langle {}^*\mathbf{d_f^i}, {}^*\mathbf{d_t^i}/\sigma \right\rangle\right)}{\mathbb{E}_i\left[exp\left(\left\langle {}^*\mathbf{d_f^i}, {}^*\mathbf{d_t^i}/\sigma \right\rangle\right)\right]}, \qquad (10)$$

where the higher textual similarity pseudo-pair has, the more learning signal the pseudo-pair needs to be provided, and vice versa. In this case, our multiview alignment is semantics-aware, which is different from existing multiview alignment methods [43, 51] that evenly align multiview data with different semantic similarities.

Video-text Modal Alignment. Until now, although the above statement provides a seemingly finalized solution in the case where the multiview pseudo-pair semantic information is not exactly the same, it is not data-efficient for every first-person sample. This is because it cannot help first-person videos in pseudo-pairs with low semantic similarities gain learning signals from others.

In order to permit all first-person videos to obtain knowledge from samples with different views or modalities, we employ video-text contrastive alignment to make the model training more data-efficient. As we mentioned before that every first-person and third-person video has an accurate textual description of their semantic information, this multimodal alignment can further help the neural network learn useful knowledge from these phrases.

Technically, given a batch of N normalized first-person and third-person features $\{\mathbf{z_f^1}, ..., \mathbf{z_f^N}\}$ and $\{\mathbf{z_t^1}, ..., \mathbf{z_t^N}\}$, and their corresponding normalized textual description vectors $\{\mathbf{d_f^1}, ..., \mathbf{d_f^N}\}$ and $\{\mathbf{d_t^1}, ..., \mathbf{d_t^N}\}$, the multimodal contrastive alignment loss (L_m) for both first-person videos and third-person videos is defined as:

$$L_m = L_{dc}^{(\mathbf{z_f}, \mathbf{d_f})} + L_{dc}^{(\mathbf{d_f}, \mathbf{z_f})} + L_{dc}^{(\mathbf{z_t}, \mathbf{d_t})} + L_{dc}^{(\mathbf{d_t}, \mathbf{z_t})}. \qquad (11)$$

3.4. Training and Evaluation Pipelines

The two proposed cross-view and cross-modal losses, along with the task-specific losses for both first-person videos (L_f) and third-person videos (L_t), are combined together during the training procedure to construct the final training loss:

$$\mathbb{L} = L_f + w_t \cdot L_t + w_{aw} \cdot L_{aw} + w_m \cdot L_m. \qquad (12)$$

Note that our model is trained in multiple stages. First, both first-person and third-person video encoders are initialized by weights pretrained on Kinetics-400 [16]. Then we train the third-person video encoder and task-specific head on the third-person video dataset with third-person video task-specific loss (L_t) only. Next, we load this trained weight for the third-person video encoder and task-specific head and do training for the whole framework with the overall loss in Equation 12. Finally, we evaluate the first-person video performance of our method using the first-person video encoder and task-specific head.

The specific task for both first-person and third-person videos we discussed in this paper is video classification. Therefore, both L_f and L_t are classification losses; w_t, w_{aw}, and w_m are the loss weights.

4. Experiments

4.1. Experiment Settings

4.1.1 Dataset

In this section, we introduce the standard benchmarks we used for evaluation in the experiments.

LEMMA [15] contains 648 third-person videos and 445 first-person videos. Among them, 136 activities were performed in kitchens, and the remaining 188 were in the living rooms. It contains 25 verb classes, 64 noun classes, and 863 action classes in their action recognition task, and includes not only single-agent but also multi-agent videos. We utilize single-agent third-person videos only as our unpaired third-person samples.

Charades-Ego [39, 40] contains 68,536 activity instances in paired 68.8 hours of first-person and third-person videos. It includes 157 action classes. We only utilize first-person videos during training and inference.

EPIC-Kitchens [4] is a large-scale egocentric video benchmark recorded in kitchen environments. It contains

55 hours of video with a total of 39.6K action segments and is labeled with 352 object and 125 verb classes.

EPIC-Kitchens-100 [5] extends EPIC-Kitchens [4] to 100 hours and 89,977 action segments. We employ this as well as EPIC-Kitchens [4] to validate our approach.

4.1.2 Baseline Methods

We compare the proposed approach with the following baseline methods, including typical multiview representation learning methods:

FPV Only is trained with first-person (FPV) video data only. Its framework is composed of first-person video encoder and first-person video classification head.

Typical CL is widely employed for view- or modal-invariant contrastive learning (CL) [43, 54], and vision-language models [32, 48, 21, 49, 24, 55]. Here we implement it with the loss function in Equation 6. Similar to our framework, videos from different views are fed to different encoders. We employ the same pseudo-pair construction method and the same inference strategy as our framework for this baseline method.

Triplet is another widely used multiview representation learning method [51, 39]. We employ the same implementation protocol following [51]. Moreover, we leverage this method to tackle our problem following the same procedures as "Typical CL".

4.1.3 Implementation Details

To ensure fairness, we keep all hyperparameters the exact same for all baselines and our method. For Charades-Ego [39, 40], we use the base learning rate of 0.25, batch size of 32 and maximum epoch of 60. For EPIC-Kitchens [4] and EPIC-Kitchens-100 [5], we use the base learning rate of 0.01, weight decay of 0.0001, batch size of 64 and maximum epoch of 30. Following [22], all models are trained using the stochastic gradient descent optimizer [1] with the momentum of 0.9. For Charades-Ego [39, 40], we use the cosine learning rate decay. For EPIC-Kitchens [4] and EPIC-Kitchens-100 [5], we use sequential learning rate decay. The backbones we use are ResNet-3D-50 [13] and SlowFast [8]. More details about the implementation of backbones will be provided in the following experimental sections.

4.2. Validity of Proposed Method

Table 1 shows the performance of different training strategies, where all methods are implemented with the ResNet-3D-50 [13] encoder. Our proposed method outperforms the "FPV only" baseline model on both Charades-Ego and EPIC-Kitchens (+1.6% mAP on Charades-Ego,

Dataset	C-Ego	EK-55 (Verb)		EK-55 (Noun)	
Evaluations	mAP (%)	Top-1 acc. (%)	Top-5 acc. (%)	Top-1 acc. (%)	Top-5 acc. (%)
FPV Only	24.7	61.1	87.1	45.8	70.4
Typical CL	25.1	60.5	86.2	45.3	68.6
Triplet	25.2	61.5	86.8	45.7	69.7
SUM-L (Ours)	**26.3**	**62.3**	**87.2**	**47.3**	**70.9**

Table 1. Comparison with baseline methods. The backbone we use is R3D-50 [13] for all methods. "C-Ego" denotes Charades-Ego [39] dataset and "EK-55" denotes EPIC-Kitchens [4] dataset.

and +1.2%/+1.5% on EPIC-Kitchens verbs/nouns top-1 accuracies). This demonstrates that under partially similar semantics situations, with our method, the third-person videos can help to improve first-person video representations.

Moreover, different from our method, previous multiview learning methods ("Typical CL" and "Triplet") cannot bring significant and consistent improvement compared with "FPV only". For example, "Typical CL" brings only 0.4% improvement on Charades-Ego (far less than ours) and even deteriorates performance on EPIC-Kitchens. Instead, our proposed method outperforms existing alignment methods up to +1.2% mAP on Charades-Ego and +1.8%/+2.0% on EPIC-Kitchens verbs/nouns top-1 accuracies. We believe this is due to the following reasons. (1) Our method semantics-awarely aligns unpaired data, instead of evenly aligning pseudo-pairs with different semantic similarities. (2) Our method filters the negative effects caused by multiview pseudo-pairs of low semantic similarity. This is different from typical multiview methods which naively align all the pairs. (3) Our method permits all first-person videos to obtain knowledge from samples with different views or modalities, which improves data efficiency.

We further verify the effectiveness of our proposed method with a larger backbone, SlowFast [8], as the encoder, and we compare with recent state-of-the-art methods on three popular egocentric action recognition benchmarks, Charades-Ego [39], EPIC-Kitchens [4], and EPIC-Kitchens-100 [5]. Note that evaluation scores of state-of-the-arts are reported directly from their paper. The experimental results employing SlowFast [8] encoder reconfirm the following findings. (1) Our method is effective for cross-view pseudo-pair alignment when there is only partially same multiview semantics. (2) Compared with existing multiview alignment methods, the improvement based on "FPV Only" of our method is still more significant and consistent. Next, we analyze the important details of the experiments.

Comparison on Charades-Ego Table 2 demonstrates the effectiveness of our method for cross-view pseudo-pair alignment when there is only partially same semantics. Note that our method and Ego-Exo [22] complement each other. Thus, we can also incorporate our method into Ego-

	Methodology	mAP (%)
state-of-the-arts	Actor [39]	20.0
	SSDA [3]	23.1
	I3D [22]	25.8
	Ego-Exo [22]	28.7
	Ego-Exo* [22]	30.1
w/o "EE"	FPV Only	27.4
	Typical CL	27.9
	Triplet	28.1
	SUM-L (Ours)	28.5
w/ "EE"	FPV Only	29.5
	Typical CL	29.1
	Triplet	28.6
	SUM-L (Ours)	**30.7**

Table 2. Comparison on Charades-Ego. Our method and all baseline methods are implemented with SlowFast-R101 [8] backbone. "EE" denotes training based on pretrained weights provided by [22].

	Methodology	Verb (%)	Noun (%)
state-of-the-arts	SlowFast-R50 [8]	64.1	48.6
	Ego-Exo [22]	66.0	49.4
	Ego-Exo* [22]	66.4	49.8
w/o "EE"	FPV Only	63.1	48.4
	Typical CL	63.9	49.6
	Triplet	63.9	48.8
	SUM-L (Ours)	65.0	49.5
w/ "EE"	FPV Only	65.8	49.5
	Typical CL	63.8	48.4
	Triplet	63.9	49.6
	SUM-L (Ours)	**66.7**	**50.2**

Table 3. Comparison on EPIC-Kitchens. The accuracies are measured by top-1 verb & noun classification accuracies. Our method and all baseline methods are implemented with SlowFast-R50 [8] backbone.

Exo [22]. As shown in Table 2, based on pretrained weight on [22], our proposed method further improves the performance of "FPV only" by 1.2% mAP, and achieves a new state-of-the-art on Charades-Ego [39].

Comparison on EPIC-Kitchens As we can see from Table 3, our method still brings clear improvement compared with "FPV Only" method (+1.9%/+1.1% verbs/nouns top-1 accuracies). Furthermore, with the help of pretrained weights from [22], our method still outperforms most state-of-the-art methodologies, which indicates the effectiveness of our method. Last but not least, our method outperforms existing alignment methods up to +2.9%/+1.8% of verbs/nouns top-1 accuracies.

Comparison on EPIC-Kitchens-100 As shown in table 4, our method also obtains competitive results on EPIC-Kitchens-100 [5] dataset when trained based on weights from [22]. It is worth noting that our method gains the first place in noun accuracy and third place in verb accu-

	Methodology	Verb (%)	Noun (%)
state-of-the-arts	TSN [45]	60.2	46.0
	TSM [23]	67.9	49.0
	TempAgg [36]	59.9	45.1
	Ego-Exo [22]	67.0	52.9
	Ego-Exo* [22]	67.5	52.9
	IPL [47]	**68.6**	51.2
w/ "EE"	SUM-L (Ours)	67.0	**53.4**

Table 4. Comparison on EPIC-Kitchens-100 with state-of-the-arts that implement CNN backbone. The accuracies are measured by top-1 verb & noun classification accuracies. We implement SlowFast-R101 [8] backbone.

L_a	L_{aw}	L_m	mAP (%)
×	×	×	24.7
×	✓	×	25.6
×	✓	✓	26.3
✓	×	✓	26.1

Table 5. Different training task combinations of our proposed method. Combining semantics-aware unpaired cross-view alignment with video-text alignment, our method clearly improves egocentric classification performance on the Charades-Ego dataset.

racy compared with state-of-the-art methodologies which employ convolutional neural network backbones.

4.3. Ablation Studies

In this section, we conduct the ablation studies on Charades-Ego [39] dataset. The backbone we employ is ResNet-3D-50 [13].

4.3.1 Impact of Different Alignment Tasks

The experiment results are shown in Table 5. We find our proposed unpaired cross-view alignment method itself brings 0.9% mAP improvement comparing with the "FPV Only" method. It denotes that the method can help learn useful knowledge from unpaired third-person view samples when the cross-view semantic information is not identical. Moreover, our proposed video-text modal alignment method further boosts first-person performance by 0.7% mAP compared with implementation with cross-view alignment only. We think this is because the proposed video-text modal alignment ensures the alignment of all first-person video samples, and further improves the semantics utilization for both first-person and third-person videos. Besides, implementing our cross-view alignment without semantic weighting causes a slight 0.2% drop, which manifests the usefulness of knowledge from the large language model. This knowledge provides a reference to measure the semantic similarity between cross-view videos.
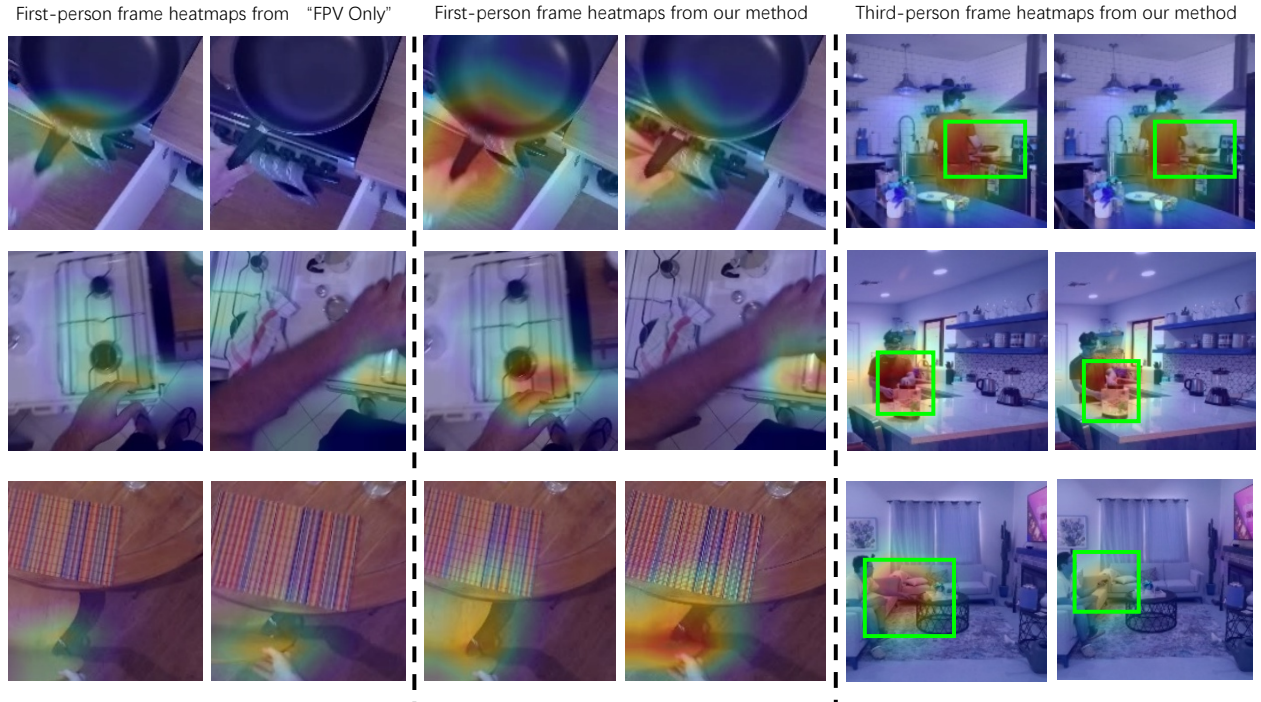
First-person frame heatmaps from "FPV Only"   First-person frame heatmaps from our method   Third-person frame heatmaps from our method

Figure 4. Comparison of class activation map visualizations between the "FPV Only" method and our method. The cross-view similar semantics for data in the first row is "holding pan". The cross-view similar semantics for data in the second and third rows is "take".

4.3.2 Impact of Third-person View Backbone

The original setup of the third-person video encoder is mentioned in Sec. 3.4. Here we study the effect of other intuitively feasible third-person video encoder implementations. The experiment results are shown in Table 6.

First, we load the pretrained weight of the third-person video and freeze it during the training process for the whole framework shown in Figure 3. We find freezing the third-person video encoder causes a 0.5% mAP drop compared with the original setup. This indicates that for the pretrained third-person video encoder, allowing training more parameters can better help first-person video learn effective information from third-person samples. Moreover, naively employing shared encoder weights for different views, which follows [2], can cause clear performance drop (1.0% mAP drop) to first-person video learning. We think this is due to the big variances between first-person and third-person samples. Finally, the same initialization (pretrained on Kinetics-400 [16]) for both first-person and third-person video encoders causes a slight 0.2% mAP drop. It indicates that for the third-person video encoder, employing existing weights of third-person samples for optimization has a positive effect on first-person video learning.

Setup	mAP (%)
Freezed TPV Backbone	25.8
Shared Multiview Encoder Weights	25.3
Same Initialization with FPV Encoder	26.1
Original	26.3

Table 6. Impact of different third-person video encoder implementations. The "Original" setup is described in Sec. 3.4.

4.4. Visualization

We visualize the first-person and third-person video class activation maps (CAM) [56] on EPIC-Kitchens [4] dataset. We employ first-person video encoder weights from "FPV Only" method (61.1% & 45.8% top-1 verb & noun accuracies) and our method (62.3% & 47.3% top-1 verb & noun accuracies). As shown in the first and second columns of Figure 4, comparing with "FPV Only" method, our method can help the model be more responsive to positive spatio-temporal locations of video clips.

Moreover, from the third column of Figure 4, we find that our model also has a strong response to the third-person video locations which have overlapped semantics with the first-person video (see bright green bounding boxes of Figure 4). It shows that our method successfully utilizes knowledge from unpaired third-person videos to help learn first-person video representations.

5. Conclusions

In this paper, we tackle a new problem in first-person and third-person video learning and propose SUM-L. It can distill knowledge from unpaired, cross-dataset third-person video without exact same multiview semantic information. Our experiments show the validity of our proposed method. Besides, under the more challenging scenario than typical paired or unpaired multimodal or multiview learning, our method outperforms existing view-alignment methods.

However, improvements from SUM-L are not very significant. One possible reason is due to the relatively small amount of third-person data. It will be interesting to explore more unpaired third-person video datasets. It will be also interesting to study more downstream tasks, such as unpaired cross-view retrieval. We leave these as future work.

References

[1] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993. 6

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4, 8

[3] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020. 7

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2, 5, 6, 8

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 1, 2, 6, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[7] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6, 7

[9] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019. 1, 2

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4

[12] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 6, 7

[14] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 754–769, 2018. 1, 2

[15] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020. 1, 2, 3, 5

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 5, 8

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 4

[18] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667, 2019. 4

[19] Jogendra Nath Kundu, Siddharth Seth, Anirudh Jamkhandi, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and

Venkatesh Babu R. Non-local latent relation distillation for self-adaptive 3d human pose estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 158–171. Curran Associates, Inc., 2021. 1, 3

[20] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015. 1, 2

[21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3, 6

[22] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 1, 2, 6, 7

[23] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 7

[24] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 6

[25] Ting Liu, Jennifer J Sun, Long Zhao, Jiaping Zhao, Liangzhe Yuan, Yuxiao Wang, Liang-Chieh Chen, Florian Schroff, and Hartwig Adam. View-invariant, occlusion-robust probabilistic embedding for human pose. *International Journal of Computer Vision*, 130(1):111–135, 2022. 3

[26] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013. 1, 2

[27] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18177–18186, June 2022. 1

[28] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. 1

[29] Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, and Sei Naito. An attention-based activity recognition for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 551–556, 2014. 1, 2, 3

[30] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 1, 2, 3

[31] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[33] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 3

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. 2, 3, 4, 5

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 3

[36] Fadime Sener, Dibyadip Chatterjee, and Angela Yao. Technical report: Temporal aggregate representations. *CoRR*, abs/2106.03152, 2021. 7

[37] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020. 2

[38] Yuge Shi, Brooks Paige, Philip Torr, and N Siddharth. Relating by contrasting: A data-efficient framework for multimodal generative models. In *International Conference on Learning Representations*, 2020. 1

[39] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018. 1, 2, 3, 5, 6, 7

[40] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 3, 5, 6

[41] Yu-Chuan Su and Kristen Grauman. Detecting engagement in egocentric video. In *European Conference on Computer Vision*, pages 454–471. Springer, 2016. 1, 2

[42] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Learning to recognize actions on objects in egocentric video with attention dictionaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2, 3, 5, 6

[44] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2, 3

[45] 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 *European conference on computer vision*, pages 20–36. Springer, 2016. 7

[46] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12249–12256, 2020. 1, 2

[47] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 7

[48] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021. 3, 6

[49] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3, 6

[50] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. *arXiv preprint arXiv:2110.06848*, 2021. 2, 4, 5

[51] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1358–1366, 2019. 1, 2, 3, 5, 6

[52] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First-and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[53] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N. Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[54] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J. Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12793–12802, June 2021. 2, 3, 6

[55] Yue Zhao, Ishan Misra, Philipp Krahenbuhl, and Rohit Girdhar. Learning video representations from large language models. *ArXiv*, abs/2212.04501, 2022. 6

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 8