First- And Third-Person Video Co-Analysis By Learning Spatial-Temporal Joint Attention

Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu, *Member, IEEE*

Abstract—Recent years have witnessed a tremendous increase of first-person videos captured by wearable devices. Such videos record information from different perspectives than the traditional third-person view, and thus show a wide range of potential usages. However, techniques for analyzing videos from different views can be fundamentally different, not to mention co-analyzing on both views to explore the shared information. In this paper, we take the challenge of cross-view video co-analysis and deliver a novel learning-based method. At the core of our method is the notion of "joint attention", indicating the shared attention regions that link the corresponding views, and eventually guide the shared representation learning across views. To this end, we propose a multi-branch deep network, which extracts cross-view joint attention and shared representation from static frames with spatial constraints, in a self-supervised and simultaneous manner. In addition, by incorporating the temporal transition model of the joint attention, we obtain spatial-temporal joint attention that can robustly capture the essential information extending through time. Our method outperforms the state-of-the-art on the standard cross-view video matching tasks on public datasets. Furthermore, we demonstrate how the learnt joint information can benefit various applications through a set of qualitative and quantitative experiments.

Index Terms—Egocentric perception, joint attention, shared representation, first-person video, third-person video, co-analysis, deep learning

1 Introduction

With the improvement of photography and popularization of cameras, a tremendous number of videos have been shot. Thus, video analysis, a technique used for automatically exploiting and comprehending the video contents, has attracted attention in both computer vision and machine learning field. This technique can be widely applied in various areas, for instance, object detection, action recognition, video surveillance, and robotics *et al*.

Among all the real-world videos, most of them are third-person videos, meaning that they are captured from a third-person viewpoint by the camera not associated with any person or object in the video. This type of video is the most common one in our daily lives. However, different from the third-person videos, the first-person videos can be captured by wearable cameras and see the visual scenes from the first-person perspective which is inherently human-centric. Although the first-person videos are still a minority, its rapid growth in quantity has attracted increasing attention from both industry and academia [1], [2], [3].

In the literature, a large number of techniques have been proposed to analyze the third-person videos, while the others focus on the first-person videos. Since the first- and third-person videos can capture the same scene from different perspectives, it is natural to further analyze them jointly. However, there exists few research investigating this problem. Very recently, Sigurdsson *et al*. [4] attempted to align video frames from first- and third-person domains by learning a joint embedding, which produces first-person video representations and third-person video representations in the same space. In this manner, the method of [4] learns a shared representation between these two viewpoints. This work has demonstrated the benefit of linking the two views. However, its performance in finding the correct corresponding regions is still unsatisfactory. To the best of our knowledge, exploring joint information from both views, although interesting and useful, remains a challenge.

The major difficulty lies in that different viewpoints produce quite different appearances, scales and locations of the same scene and target, *e.g*., the visual area in the first-person image only corresponds to a small and deformed part of that in the third-person view. Therefore, it is ineffective to learn the shared representation of the two views' frames directly as in [4]. Instead, we need to model and extract a more fundamental relationship between the two views and then use it to guide the shared representation learning. This enlightens our idea in this research.

Following the above discussion, in this paper, we propose to learn the shared representation from the first-person and third-person views more effectively and robustly. Our key idea is to learn the "joint attention" across different views (as shown in Fig. 1), which plays an important role in the extraction of shared representation between the two

- *Huangyue Yu, and Yunfei Liu are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China. E-mail: {yuhuangyue, lyunfei}@buaa.edu.cn.*
- *Minjie Cai is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. E-mail: caiminjie@hnu.edu.cn.*
- *Feng Lu is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, also with the Peng Cheng Laboratory, Shenzhen 518055, China, and also with the Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing 100191, China. E-mail: lufeng@buaa.edu.cn.*

Manuscript received 21 February 2020; revised 19 August 2020; accepted 5 October 2020. Date of publication 12 October 2020; date of current version 5 May 2023.
(Corresponding Author: Feng Lu.)
Recommended for acceptance by A. Furnari, D. Crandall, D. Damen, K. Grauman, and G.M. Farinella.
Digital Object Identifier no. 10.1109/TPAMI.2020.3030048

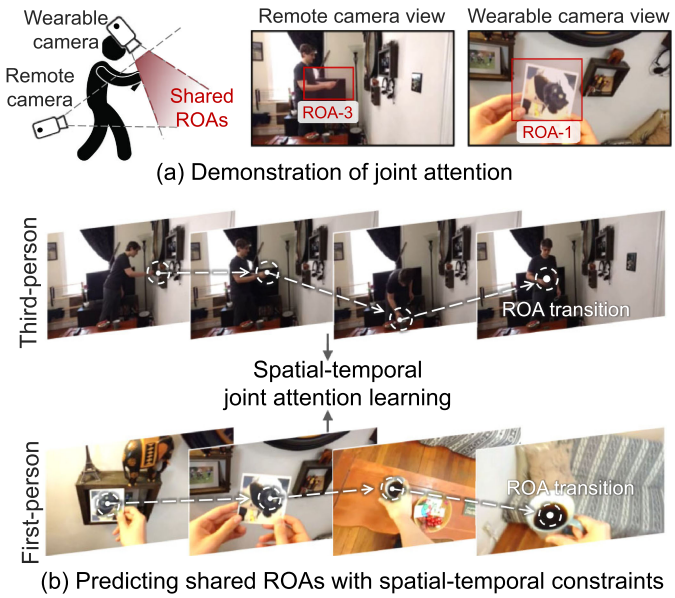


Fig. 1. The motivation of this work. (a) First- and third-person videos are captured simultaneously by a wearable camera and a remote camera. Joint attentions are defined as the corresponding attention regions (Shared Regions of Attentions, Shared ROAs) across two viewpoints, where ROA-3 indicates the ROA in third-person video and ROA-1 indicates that in first-person video. (b) The goal of this work is to find the joint attention regions automatically from two views by utilizing the spatial-temporal constraints on ROAs. The joint attention helps extract shared representations between two views, which benefits a range of first/third/mixed view-based applications.

views. To be more specific, we define the region of attention (ROA) in both the first- and the third-person views, and make the assumption that the shared representation, which effectively associates the knowledge from the two views, should be extracted based on the joint attention regions (shared ROAs). In this manner, our shared representation can be learned on the basis of physically meaningful shared ROAs rather than the original cross-view video frames.

In order to extract the joint attention regions from two views and learn the shared representation subsequently, we propose a novel learning framework based on Convolutional Neural Networks (CNNs), namely Joint Attention Network (JANet). It uses video frames from the first- and third-person views as input and incorporates self-supervised attention learning to extract shared ROAs without explicit annotations. In particular, channel attention vectors of the two views' frames are first generated, which are enforced to be similar to obtain the shared ROAs. Finally, shared representations between the two views are learned via a triplet loss to enforce that the feature representations extracted from shared ROAs of corresponding frames are close to each other, and vice versa.

Extended from the JANet, we further propose the Temporal Joint Attention Network (T-JANet) with extra supervision of temporal information. This is considered a major difference and improvement compared to an early conference version of this work [5]. The T-JANet models and utilizes the temporal transition of ROAs in the video as a constraint to learn a spatial-temporal joint attention, which shows a significant improvement in accuracy. In particular, after extracting the channel attention vectors from two views, a specific temporal constraint is proposed as the extension of self-supervised attention learning module to handle the temporal transition of shared ROAs. In this manner, the T-JANet becomes capable of modeling and learning the joint attention temporally. By fusing the final features through time, the method obtains the shared representations under both spatial and temporal constraints.

The advantages of our methods have been demonstrated via the standard cross-view video matching tasks in [4]. In addition, we apply the joint attention as well as the shared representation learned by our methods to various applications, including gaze prediction, cross-view image co-segmentation, video summarization, and zero-shot action recognition. We show promising results both quantitatively and qualitatively in these applications, which indicates that our method is not only effective but also versatile. The extracted joint attention and shared representation can be flexibly applied to handle different tasks to boost the performance.

Overall, our major contributions are as follows:

- We introduce the joint attention, based on which the shared representation between the first- and third-person videos can be learned more effectively and robustly.
- We propose a new multi-branch deep network for joint attention learning and shared representation learning, based on a self-supervised attention learning architecture. The network extracts important features from cross-view videos effectively.
- We model the temporal transition of ROAs by learning with temporal constraint, based on which the joint attention can be explored spatially and temporally.
- Comprehensive experiments on large-scale benchmark datasets for two cross-view video matching tasks demonstrate the effectiveness of our proposed method as well as the key contributions in the proposed framework.
- Additional experiments are also presented to show that our method can be used as a fundamental tool to benefit various related applications.

2 Related Works

2.1 Co-analysis of First- and Third-Person Videos

Recent studies of modeling between first- and third-person videos have been often conducted on paired videos of these two domains [4], [6], [7], [8], [9], [10]. Yonetani *et al.* [6] proposed a novel face detection approach by matching camera and head motion of the same person from first- and third-person perspectives. Ardeshir and Borji *et al.* [7] matched a set of first-person videos to a set of characters (first-person camera wearers) in a top-view surveillance video using graph matching. Fan *et al.* [8] studied a similar problem by learning a joint feature embedding space from first- and third-person videos with a two-stream semi-Siamese network. Unlike [8], which requires bounding boxes ground-truth, Xu *et al.* [9] simultaneously segmented and matched the first-person camera wearers in third-person videos. Moreover, in recent years, some large-scale datasets, such as EGTEA Gaze+ [11], EgoSum+gaze [12], Charades-Ego [4] and EPIC-Kitchens [13], have been proposed to address tasks related to first-person videos.

While the above works assumed that the paired first- and third-person videos were synchronized, we propose an approach to temporally match individual video frames between these two domains. The most related work to ours is that of Sigurdsson *et al.* [4], which learned a joint embedding space between matched first- and third-person video frames. Different from [4], we propose to learn joint attention in both videos for more accurate matching.

2.2 Attention Models for Image and Video Analysis

Detecting and understanding the attention regions in images and videos have been an emerging research field in computer vision and multimedia processing these years. Attention model has shown its efficiency in various vision tasks such as person re-identification [14], [15], [16], [17], image captioning [18], [19], [20], pose estimation [21], [22], [23] and image classification [24], [25]. In 1998, Itti *et al.* [26] constructed a primary visual attention model using a bottom-up architecture. After that, various attention models with different architectures are inspired: Wang *et al.* [27] proposed a non-local blocks operation, which is related to the self-attention method. In [28], an interaction-aware attention network was presented to construct a spatial feature pyramid for obtaining more accurate attention maps by multi-scale information. Woo *et al.* [29] made an extension of squeeze-and-excitation module [30], and presented a Convolutional Block Attention Module (CBAM). Jun *et al.* [31] introduced visual attention into image segmentation, which then achieved better performance through their long-range context relationships.

Different from the previous works, we use data from different viewpoints to jointly learn the shared ROAs from videos. The shared ROAs will primarily focus on the joint attention regions from different viewpoints.

2.3 Representation Learning Among Viewpoints

Existing research on cross-view representation learning usually adopts deep metric learning with siamese (triplet) architectures [32], [33] or proposes an encoder-decoder framework with generative adversarial networks [34], [35]. Regmi *et al.* [34] addressed the novel problems of cross-view image synthesis, aerial-to-ground view and vice versa, by using conditional generative adversarial networks to learn shared representation. Hu *et al.* [32] used a triplet loss incorporating a CVM-Net for ground-to-aerial geolocalization.

In this paper, the notion of "joint attention" is to be developed, which holds the view that the shared representation among cross-views should correspond to the joint attention regions. We construct a self-supervised attention learning architecture to extract the joint attention regions from cross-views.

2.4 Video Analysis Using Temporal Information

Temporal analysis is capable of examining and modeling the behavior over time. Jeffrey *et al.* [36] conceived and first proposed a Simple Reccurent Network (SRN), which was a specific version of the back propagation neural network and accessible to process of sequential input and output. SRN is groundbreaking for many cognitive scientists and psycholinguists, since it has been particularly useful in time series prediction, such as language understanding. Long Short Term Memory (LSTM) network, proposed by Sepp Hochreiter *et al.* [37], was an extended Recurrent Neural Network (RNN). Its unique design enables LSTM to handle and predict the important issues with long intervals and even delay in the time sequence. It has been widely used in handwriting recognition, speech recognition and machine translation. In the same year, Bidirectional RNN (BRNN) was proposed by Mike *et al.* [38]. It could be trained simultaneously in positive and negative time directions. Today, the notion of "sequence to sequence" [39] based on encode-decoder structure impacts significantly on the field of machine translation and machine understanding. Many research works have been devoted to modeling the temporal structure in various applications [40], [41], [42], [43], [44], [45].

In general, the above methods directly operated on frames with RNNs. It is still non-trivial for these methods to learn from the informative regions despite that the videos contain much redundant and irrelevant information. Our method differs from these end-to-end temporal networks in learning the transition of ROA, which enables efficient and accurate learning.

3 Key Idea: Joint Attention Guided Representation Learning

The key idea of this work is the learning of joint attention across different views, which plays an important role in the extraction of shared feature representation for the first- and third-person video co-analysis. In this section, we analyze the challenges in cross-view video co-analysis and introduce the motivation behind our proposed method.

3.1 Joint Attention in Shared Representation Learning

Given videos of both first- and third-person viewpoints, our goal is to learn a shared representation that benefits video (co-) analysis and allows knowledge transfer or knowledge fusion for various tasks. This can be done by embedding visual data from two videos in a shared space, which however is quite complicated. The fundamental difference in viewpoints makes the shared representation learning much more problematic.

The exemplified case in Fig. 1b illustrates that a third-person video captures the full view of a person while a first-person video only focuses on the local region centering around the person's hand. We observe that the corresponding regions in the two views, *i.e.*, the hands and the object being manipulated, occupy a large area in the first-person view but a very small area in the third-person view. In addition, due to different viewpoints, the appearances of these regions can vary significantly. This is, in face, a common case in most first- and third-person video data. As a result of such serious misalignment between the corresponding image regions, the embedment for shared representation among views becomes more challenging.

To address such difficulty, we propose to locate corresponding regions across views. To this end, we introduce a novel concept of *joint attention*. This can be illustrated in Fig. 2a, where joint attention can be denoted by the shared Regions Of Attention (ROAs) that align the same regions in both views with different scales and orientations. The

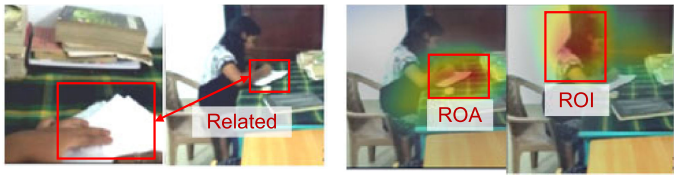


Fig. 2. Illustration of the proposed joint attention in the form of shared Regions Of Attention (ROAs). (a) The shared ROAs can locate the same hand regions in both the first- and third-person video frames. This is of great importance for the subsequent shared representation learning. (b) The shared ROAs can be quite different from traditional ROI (Region Of Interest) since they must be in consensus across views.

shared ROAs are expected to capture the most informative regions across views. The shared ROAs are in consensus across views, different from the traditional ROI (Region Of Interest) considering only the visually interesting region in a single view (Fig. 2b).

The joint attention introduced above is the core idea of our method. It can locate the corresponding regions that reflect the same and most important area across views, thus solving the misalignment problem and allowing shared representation learning. Technical details are provided in the "Attention-guided feature extraction" part of Section 4.2.

3.2 Self-Supervised Joint Attention Learning

Although the incorporation of attention mechanism might help obtain the shared representation between first- and third-person videos, identifying the regions of attention is a nontrivial task. Traditional saliency-based models tend to locate the region of interest (ROI) that is attractive to human perception [46] but not necessarily corresponds to the informative region (ROAs) shared by the two views.

In order to identify the corresponding regions, we propose a self-supervised learning technique that explores the ROAs in different views without using explicit ground truth data. Our assumption is that, the appearances of the shared ROAs, although may vary in scale and orientation, correspond to the same target or event in the scene. Following this assumption, the localization of ROAs can be well constrained by enforcing a semantic consistency between the candidate ROAs in the two views. Technical details of self-supervised learning are provided in the "Self-supervised joint attention learning" part of Section 4.2.

3.3 Simultaneous Learning of Joint Attention and Shared Representation

As explained in Section 3.1, joint attention is devoted to learning the shared representation between the first- and third-person videos. Meanwhile, the learned shared representation provides high-level features for the shared ROAs, which can naturally ensure the semantic consistency constraint for the self-supervised learning of joint attention, as explained in Section 3.2. In other words, the learning of joint attention and shared representation mutually benefit.

Therefore, we propose a simultaneous learning strategy for obtaining joint attention and shared representation. The following two constraints are applied: 1) the extracted ROAs are semantically consistent between corresponding videos, and 2) feature representations are close between corresponding videos and distant between non-corresponding videos. Technical details of simultaneous learning are provided in subsection "Shared representation learning" of Section 4.2.

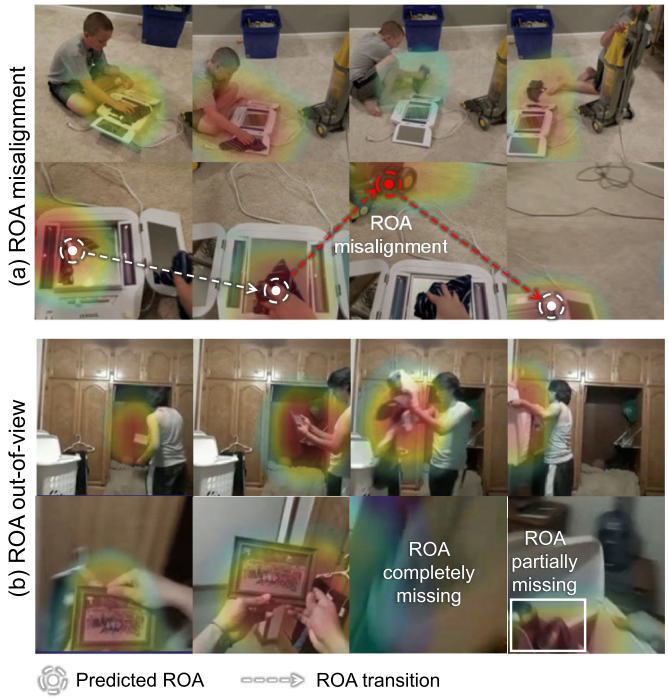


Fig. 3. Temporal constraints are proposed to solve the following problems. (a) ROA misalignment: the ROA of the third frame of the first-person video is inconsistent with its adjacent frames. (b) ROA out-of-view: the ROAs of the third/fourth frames of the first-person video are completely/partially missing.

3.4 Learning With Temporal Constraint

By now, we have described our strategy of learning joint attention and shared representation simultaneously in the first- and third-person videos. However, the current strategy only processes video frames at certain time-points separately. Consequently, unexpected failure might appear in the following cases.

- *ROA Misalignment*. The separately predicted ROAs may not be consistently accurate. As shown in Fig. 3a, the joint attention region should fall on the toy area. However, in the third frame, the ROA predicted for the first-person view has changed to a different area with similar appearance, which is even more obvious by checking its previous/next frames. This causes ROA misalignment temporarily.
- *ROA Out-of-View*. An individual frame may not contain the correct ROA when it is occasionally out of view. As a result, the desired ROA can be partially or entirely missing in the frame. As shown in Fig. 3b, the ROA is completely missing in the third frame of the first-person view, and is partially missing in the fourth frame.

Neither of the problems, which are in fact related to temporal consistency and information loss, can be easily

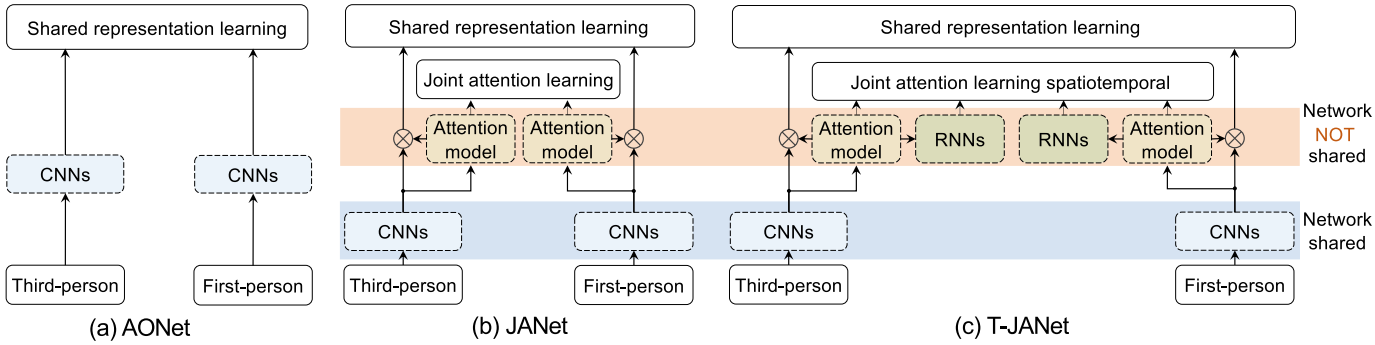


Fig. 4. Comparison of different shared representation learning strategies. (a) AONet [4] learns the shared representation by extracting features directly from CNNs. (b) Our JANet extracts joint attention to guide the shared representation learning. The two learning processes benefit each other simultaneously. (c) Our T-JANet further adopts RNNs to learn the joint attention/shared representation with temporal constraint.

addressed only in the spatial domain. To solve these problems, we propose to further utilize the temporal information which is adherently provided by the videos. By using constraints with previous and subsequent frames, we are able to ensure temporal consistency and handle the information loss to a certain degree. Technical details of incorporating temporal constraints in our method are given in Section 4.4.

4 Methodology

4.1 Overview

In this paper, we aim to learn a shared representation for co-analyzing the first- and third-person videos. In contrast with the previous works that learned a shared representation directly from videos, we incorporate attention mechanism into the representation learning framework. We propose a self-supervised joint attention learning module to predict spatial-temporal joint attention regions for shared representation learning, based on the assumption that shared representation should correspond to the joint attention regions from videos of different viewpoints.

In particular, [4] proposed an ActorObserverNet (AONet) to learn the cross-view shared representation by comparing the videos from two perspectives directly. The solution pipeline is shown in Fig. 4a. Inspired by [4], two shared representation learning strategies are proposed in this paper, which are summarized as follows:

- *JointAttentionNet (JANet)*. Demonstrated in Fig. 4b. At first, basic features are generated from video frames of different viewpoints by a standard CNN. Then, instead of directly comparing the CNN-based features from two viewpoints, a novel joint attention learning module is proposed to predict ROA for each viewpoint. Finally, the predicted ROAs are used as guidance to filter the CNN-based features for learning shared representation.
- *Temporal-JointAttentionNet (T-JANet)*. Demonstrated in Fig. 4c. As the extension of JANet, we adopt RNNs to explore temporal transition of joint attention and learn share representation temporally.

4.2 Architecture of JANet

The architecture of our proposed JANet is presented in Fig. 5, which is mainly composed by three modules:

- *Attention-guided feature extraction module*. Designed for *shared representation learning* (Section 3.1).
- *Self-supervised joint attention learning module*. Designed for *joint attention learning* (Section 3.2).
- *Shared representation learning module*. Designed for *simultaneous learning* (Section 3.3).

Attention-Guided Feature Extraction. This module focuses on extracting discriminative features for shared representation learning, which is guided by the learned attention. In brief, the module consists of a backbone CNN model for extracting feature maps (F_c^x, F_c^y, F_c^z) for a triple of input frames (x, y, z) and an attention model for generating channel attention vectors which represent relative importance of different channels (i.e., semantic information) of the feature maps.

To generate channel attention vectors, we follow the same setting of the channel attention module in [29] and convolutional block attention module [30]. The network for generating channel attention vectors is composed by two global pooling layers and one multi-layer perceptron (MLP). It takes feature maps (F_c^x, F_c^y, F_c^z) from the backbone CNN model as input and outputs 1D channel attention vectors $(\mathbf{M}_c^x, \mathbf{M}_c^y, \mathbf{M}_c^z)$, where $F_c \in \mathbb{R}^{c\times h\times w}$, $\mathbf{M_c} \in \mathbb{R}^{c\times 1\times 1}$, c denotes the number of channels and (h, w) denotes the spatial size of feature maps. The original feature maps (F_c^x, F_c^y, F_c^z) are then augmented by the generated channel attention vectors and the final feature representations $(F_c^{x\prime}, F_c^{y\prime}, F_c^{z\prime})$ are defined as $F_c' = \mathbf{M}_c(F_c) \otimes F_c$, where $\otimes$ denotes element-wise multiplication. Noted that while the parameters of the backbone CNN model for extracting feature maps are shared by different viewpoints, the parameters of the attention model are learned separately. This is because 1) the video appearance is different for different views and 2) it is difficult to estimate attention with the same network.

Self-Supervised Joint Attention Learning. In this module, we aim to learn joint attention regions between first- and third-person videos which focus on the same action recorded from two viewpoints. We develop a self-supervised learning approach by enforcing the semantic consistency between ROAs of corresponding video frames. The inputs are the pairs of channel attention vectors $(\mathbf{M}_c^x, \mathbf{M}_c^y)$ generated from a first-person video frame and the corresponding third-person video frame. We compare the two vectors with a L2-based distance metric which enforces similarity

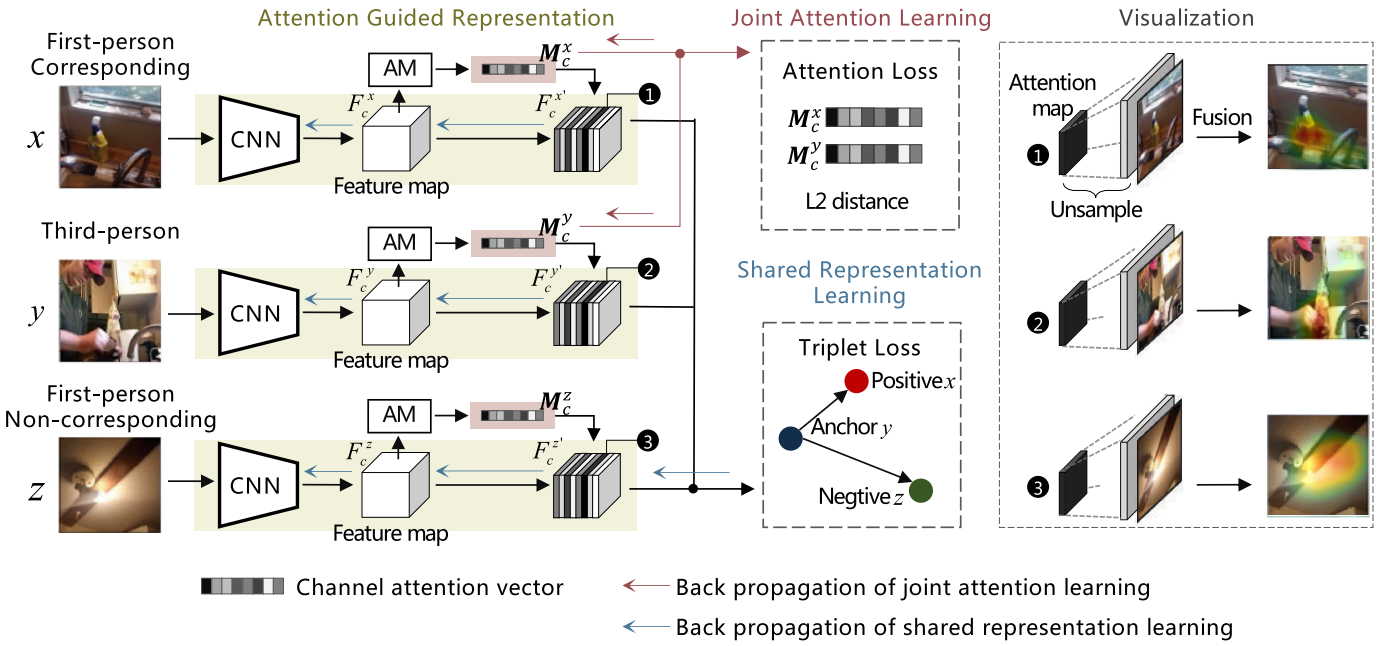


Fig. 5. The architecture of our JANet. Given a triple of frames (x, y, z) as input, the attention-guided representation module is developed to extract channel attention vectors $(\mathbf{M}_c^x, \mathbf{M}_c^y, \mathbf{M}_c^z)$ and feature representations $(F_c^{x\prime}, F_c^{y\prime}, F_c^{z\prime})$ from the intermediate feature maps (F_c^x, F_c^y, F_c^z). The joint attention learning module encourages similarity of channel attention vectors between a third-person frame $\mathbf{M}_c^y$ and the corresponding first-person frame $\mathbf{M}_c^x$. The shared representation learning module explores the common information between two viewpoints to obtain the shared representation. The attention model is represented by "AM". We conduct the weighted average on feature maps F_c based on the channel attention vectors $\mathbf{M_c}$ to generate spatial attention maps. The visualizations are obtained by overlapping the spatial attention maps with original frames and are shown on the right.

between channel attention vectors of corresponding frames and thus guarantees the semantic consistency of ROAs between first- and third-person videos.

Shared Representation Learning. In this module, we learn shared representation between first- and third-person videos based on the feature representations $(F_c^{x\prime}, F_c^{y\prime}, F_c^{z\prime})$ that are filtered by the predicted attention at the previous stage. We adopt a triplet loss to enforce that the feature representations of corresponding frames (x and y) are close to each other, and vice versa.

4.3 Loss Function of JANet

Here we describe the objective function used to train JANet. The objective function consists of two loss functions:

- *Attention loss*. It is designed for *joint attention learning* (Section 3.2).
- *Triplet loss*. It is designed for *simultaneous learning* (Section 3.3).

Attention loss is denoted as $\mathcal{L}_{\mathcal{AL}}(x, y)$, which enforces similarity between channel attention vectors of the corresponding first- and third-person video frames. It is formulated as:

$$\mathcal{L}_{\mathcal{AL}}(x, y) = \left\|\mathbf{M}_c^x - \mathbf{M}_c^y\right\|_2, \tag{1}$$

where $\|\cdot\|_2$ denotes the L2 norm.

Triplet loss is denoted as $\mathcal{L}_{\mathcal{TL}}(x, y, z)$, which enforces similarity between corresponding feature representations $F_c^{x\prime}$ and $F_c^{y\prime}$, and penalizes similarity between non-corresponding feature representations $F_c^{y\prime}$ and $F_c^{z\prime}$. It is formulated as:

$$\mathcal{L}_{\mathcal{TL}}(x, y, z) = \frac{e^{\left\|F_c^{x\prime} - F_c^{y\prime}\right\|_2}}{e^{\left\|F_c^{x\prime} - F_c^{y\prime}\right\|_2} - e^{\left\|F_c^{y\prime} - F_c^{z\prime}\right\|_2}}. \tag{2}$$

Following [4], we compute a normalized weight for all sampled frames of the same video in order to assign importance weight $w(x, y, z)$ for each triplet. The final loss $\mathcal{L}(x, y, z)$ is composed as:

$$\mathcal{L}(x, y, z) = [\mathcal{L}_{\mathcal{TL}}(x, y, z) + \lambda \cdot \mathcal{L}_{\mathcal{AL}}(x, y)] \cdot w(x, y, z), \tag{3}$$

where λ is a hyper parameter used to balance the relative contributions of different losses. Since $\mathcal{L}(x, y, z)$ is weighted by $w(x, y, z)$, the optimization is simply a weighted version of the original back-propagation. The intuition of importance weight is to decrease the negative impact of uninformative frames that are recorded under unstable conditions. More technical details about the procedure of computing $w(x, y, z)$ can be found in Section 3.3 of [4]. We empirically set $\lambda = 2.5$ in our experiments.

4.4 Architecture of T-JANet

The key difference between JANet and T-JANet is that JANet only considers the spatial information of videos, while T-JANet takes both spatial and temporal information into consideration. We adopt LSTM architecture (L) for shared representation learning with temporal constraint (Section 3.4). The details of this extended component are presented in Fig. 6.

Different from JANet, the T-JANet takes a triple of videos (X, Y, Z) as input, and

(x_t, y_t, z_t) represent the triple of video frames at time t from (X, Y, Z). We extract channel attention vectors following the same setting of JANet. $(\mathbf{M}_c^{x_t}, \mathbf{M}_c^{y_t}, \mathbf{M}_c^{z_t})$ denote the channel attention vectors from (x_t, y_t, z_t).

A LSTM network is adopted to capture the temporal information of channel attention vectors. The input of LSTM at time t is the averaged value of channel attention vectors at

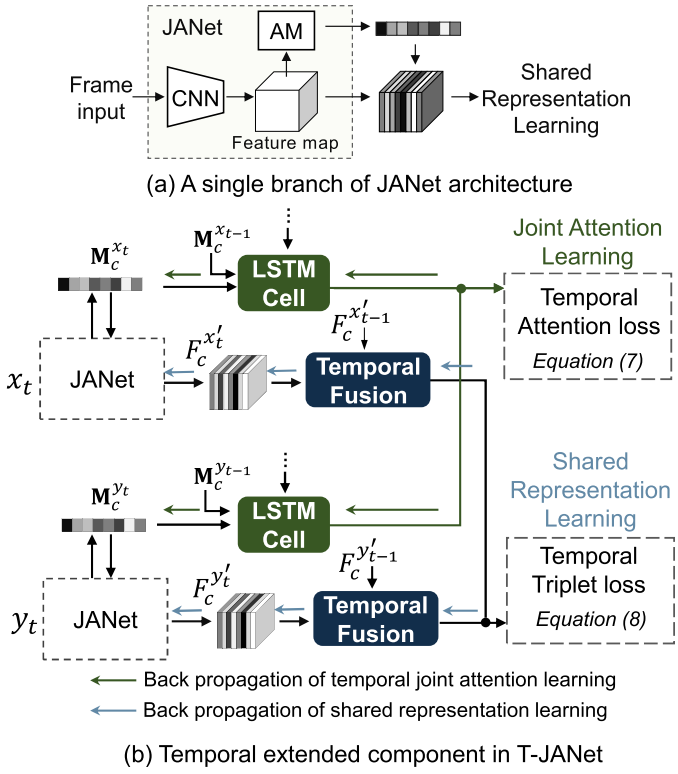


Fig. 6. Illustration of the temporally extended component in T-JANet. (a) We extract channel attention vectors $(\mathbf{M}_c^{x_t}, \mathbf{M}_c^{y_t})$ and feature representations following the same setting of JANet. (b) A LSTM network is adopted to capture the temporal information of channel attention vectors. Temporal attention loss prompts similarity of the LSTM outputs between third-person frame and corresponding first-person frame to learn the joint attention temporally. Temporal fusion is adopted to incorporate feature representations from previous frame.

current time t and previous time $t-1$. The equation for generating the LSTM output is (take input video X for example):

$$h_t^x = L[h_{t-1}^x, A(\mathbf{M}_c^{x_t}, \mathbf{M}_c^{x_{t-1}})], \tag{4}$$

where $A(\cdot)$ indicates the averaging operation.

4.5 Loss Function of T-JANet

The objective function of T-JANet consists of two loss functions:

- *Spatiotemporal attention loss*. It is designed for *joint attention learning* (Section 3.2) and *temporal constraint* (Section 3.4).
- *Temporal triplet loss*. It is designed for *simultaneous learning* (Section 3.3).

Spatiotemporal attention loss is denoted as $\mathcal{L}_{\mathcal{STAL}}(\mathrm{X},\mathrm{Y})$ which enforces joint attention between corresponding first- and third-person videos spatially and temporally. The spatial attention loss $\mathcal{L}_{\mathcal{S}-\mathcal{AL}}(\mathrm{X},\mathrm{Y})$ follows the same construction as Equation (1). The temporal attention loss $\mathcal{L}_{\mathcal{T}-\mathcal{AL}}(\mathrm{X},\mathrm{Y})$, based on L2 distance metric, is constructed to enforce similarity between LSTM outputs of corresponding first- and third-person videos. The spatiotemporal attention loss at time t is composed as:

$$\mathcal{L}_{\mathcal{STAL}}(x_t, y_t) = \mathcal{L}_{\mathcal{S}-\mathcal{AL}}(x_t, y_t) + \lambda_1 \cdot \mathcal{L}_{\mathcal{T}-\mathcal{AL}}(x_t, y_t), \tag{5}$$

where λ_1 is a hyper parameter used to balance the relative contributions of spatial-temporal losses. We empirically set $\lambda_1 = 1.2$ in our experiments.

Specifically, the spatial attention loss $\mathcal{L}_{\mathcal{S}-\mathcal{AL}}(\mathrm{X},\mathrm{Y})$ and the temporal attention loss $\mathcal{L}_{\mathcal{T}-\mathcal{AL}}(\mathrm{X},\mathrm{Y})$ are formulated as:

$$\mathcal{L}_{\mathcal{S}-\mathcal{AL}}(x_t, y_t) = \left\| \mathbf{M}_c^{x_t} - \mathbf{M}_c^{y_t} \right\|_2, \tag{6}$$

$$\mathcal{L}_{\mathcal{T}-\mathcal{AL}}(x_t, y_t) = \left\| h_t^x - h_t^y \right\|_2. \tag{7}$$

Temporal triplet loss is denoted as $\mathcal{L}_{\mathcal{TTL}}(\mathrm{X},\mathrm{Y},\mathrm{Z})$, which is an extended version of Equation 2 by taking the temporal information into consideration. It is formulated as:

$$\mathcal{L}_{\mathcal{TTL}}(x_t, y_t, z_t) = \frac{e^{\left\| F_c^{x'_{t,t-1}} - F_c^{y'_{t,t-1}} \right\|_2}}{e^{\left\| F_c^{x'_{t,t-1}} - F_c^{y'_{t,t-1}} \right\|_2} - e^{\left\| F_c^{y'_{t,t-1}} - F_c^{z'_{t,t-1}} \right\|_2}}, \tag{8}$$

where $(F_c^{x'_{t,t-1}}, F_c^{y'_{t,t-1}}, F_c^{z'_{t,t-1}})$ are the averaged values of the feature representations of current frame t and previous frame $t-1$ from videos $(\mathrm{X},\mathrm{Y},\mathrm{Z})$.

The final loss $\mathcal{L}(\mathrm{X},\mathrm{Y},\mathrm{Z})$ is composed as:

$$\begin{aligned}\mathcal{L}(\mathrm{X},\mathrm{Y},\mathrm{Z}) = \frac{1}{T}\sum_{t=1}^{T}\{[\mathcal{L}_{\mathcal{TTL}}(x_t, y_t, z_t) \\ + \lambda \cdot \mathcal{L}_{\mathcal{STAL}}(x_t, y_t)] \cdot w(x_t, y_t, z_t)\}.\end{aligned} \tag{9}$$

4.6 Implementation Details

Our framework is implemented by using PyTorch [47]. We apply a ResNet-152 architecture [48] as the basic CNN model. Input frames are cropped into 224×224. The first four convolutional layers of ResNet-152 are used to extract feature maps. Following the same settings of CBAM [29] and [30], we implement our channel attention module in our experiment. To reduce parameter overhead, the hidden activation size is $\mathbb{R}^{c/r \times 1 \times 1}$ and the reduction radio r is set as 8. The channels of feature maps and the dimension of channel attention vectors are both 2048. The spatial size of feature maps as well as that of attention maps is 7×7. The triplet importance weight is learned from one fully connected layer based on the fc7 features of ResNet-152. SGD is used to train the whole model, with the learning rate of 3e-5 and batch size of 4. For T-JANet, we uniformly sampled the frames from the video with a temporal stride of τ, i.e., it processes only one out of τ frames. The value of τ is set as 30. The LSTM is a 1-layer LSTM which takes the input of (T, b, c), where T is the raw sampled length, b is batch size, and c is the dimension of channel attention vector. The dimension of hidden state of LSTM is 1024.

5 Experiment

In this section, we evaluate our method on a public dataset with pairs of first- and third-person videos. We first compare the performance of state-of-the-art method and conduct ablation study to verify the effectiveness of our proposed method (JANet and T-JANet) qualitatively and

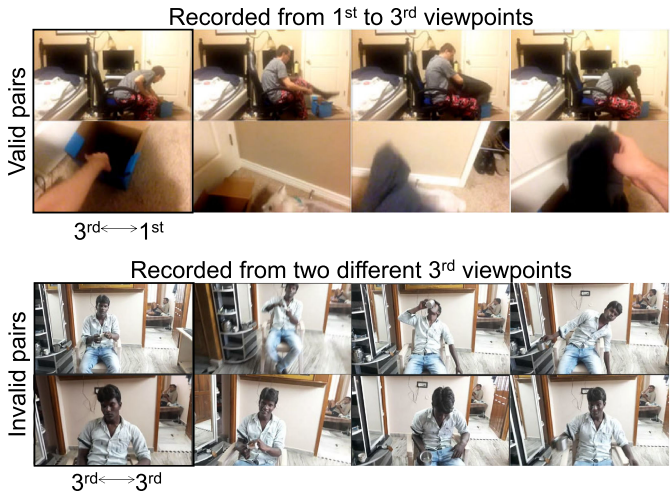


Fig. 7. Examples of video pairs from Charades-Ego dataset. The example on the top demonstrates a video pairs which are recorded from both third- and first-person viewpoints. The example on the bottom demonstrates a video pairs which are recorded from two slightly different third-person viewpoints. Our aim is to learn shared representation between first- and third-person viewpoints. Thus the example on the bottom is invalid for our research.

quantitatively. Moreover, by learning joint attention and shared representation between two viewpoints, our method can benefit various applications. Therefore, we also conduct additional experiments to demonstrate the benefits of our work on various computer vision applications.

5.1 Dataset and Evaluation Tasks

5.1.1 Dataset

To evaluate our method, we conduct experiments on the public Charades-Ego dataset [4], which consists of 4000 paired first- and third-person videos of daily indoor activities recorded by 112 persons. Since our goal is to learn a shared representation between first- and third-person videos, we focus on the video pairs that record the same action from the first- and third-person viewpoints respectively. However, there are some video pairs in the raw dataset whose viewpoints are invalid for our research. Fig. 7 demonstrates two examples from dataset. The top one displays a video pairs from first- and third-person viewpoints. The bottom one displays a video pairs recorded from two different third-person viewpoints, which are invalid for our research. Therefore, we carefully examined the dataset and removed 189 invalid video pairs in total.

5.1.2 Evaluation tasks

We evaluate the proposed method with two kinds of evaluation tasks: algorithmic evaluation and representative application evaluation.

Algorithmic Evaluation. To evaluate the algorithmic performance of our method, we consider two cross-view video matching tasks following the procedures in [4]:

- *Pairs Discrimination*. The aim of pairs discrimination is to discriminate corresponding first- and third-person image pairs from the non-corresponding ones. Classification accuracy is used as evaluation metric.
- *Best-Match Moment Localization*. The aim of best-match moment localization is to find the corresponding moment (1 second video clip) in a video, given another video moment from the other viewpoint without the knowledge of time stamp. Specifically, we compute the average of L2 distance between F_c' features of all corresponding frames in two moments as the distance of two moments. The two moments with the lowest distance are selected as the best-match moments. For evaluation, we assume that the ground truth best-match moments can be approximately obtained by temporally scaling the first-person video to have the same duration as the third-person video. Alignment error is used as the evaluation metric.

Representative Application Evaluation. The aim of representative application evaluation is to show how the proposed method could benefit various computer vision applications. We consider four related applications: gaze prediction, image co-segmentation, video summarization, and zero-shot action recognition. The details of evaluation on these applications are given in Section 5.5.

Overall, our method provides an effective way to learn the shared representation between first- and third-person videos based on the key idea of "joint attention". In addition to algorithmic evaluation on two cross-view video matching tasks (Sections 5.2, 5.3, 5.4), experiments are also conducted to show how the shared representation and joint attention which are learned through the proposed method, can help improve the performance of various related applications. By evaluation on four representative applications, we aim to show that the proposed method can be used as a fundamental tool to promote the research of first-person vision with multi-view information.

5.2 Experimental Setting for Algorithmic Evaluation

Joint Attention Network (JANet). To evaluate how different parts of JANet contribute to the final performance on the two tasks, we conduct ablation study by removing or replacing a subset of models. Details of different baselines are described as follows:

- *Without Self-Supervised Joint Attention Learning* (*Without SA*). To illustrate the contribution of the self-supervised joint attention learning in our model, we remove this part and re-train the remaining model.
- *Self-Supervised Attention Learning With CNN-Based Features*. To examine the contribution of attention information from either first- or third-person video in joint attention learning, we replace the channel attention vectors of first-person video and third-person video with original CNN-based feature maps respectively, denoted as *CNN SA /1* and *CNN SA /3*. Then the attention is learned by minimizing the L2 distance between average pooling of the feature maps and channel attention vector of the corresponding third-person video (or first-person video).
- *Triplet Loss With CNN-Based Features*. To examine the contribution of attention information from either first- or third-person video in shared representation learning, we use CNN-based features of first-person

video or third-person video to calculate triplet loss respectively. These two baselines are denoted as *CNN TL /1* and *CNN TL /3*.

- *Triplet Weight Based on Low-Level Image Features* (*Low-level TW*). We use low-level image features (the average gradient of images in our experiment) instead of high-level CNN-based features to estimate triplet weight.

Temporal Joint Attention Network (T-JANet). To evaluate the effectiveness of different parts of T-JANet, we also conduct ablation study with the following baselines:

- *Temporal Triplet Loss Without Average Value* (*TTL Without Avg*). We use the feature representation of frame t instead of the average value to calculate temporal triplet loss.
- *Self-Supervised Attention Learning Without Temporal Attention* (*Without T-AL*). To demonstrate the effectiveness of temporal information in joint attention learning, we remove the temporal attention loss from spatiotemporal attention loss and re-train the remaining model. The final loss $\mathcal{L}(\mathrm{X},\mathrm{Y},\mathrm{Z})$ is then comprised of $\mathcal{L}_{TTL}(\mathrm{X},\mathrm{Y},\mathrm{Z})$ and $\mathcal{L}_{S-AL}(\mathrm{X},\mathrm{Y})$.
- *Temporal Attention Loss Without Average Value* (*T-AL Without Avg*). We use channel attention vector of current frame t as the input of the LSTM state h_t. The equation (take input video X for example) of generating the LSTM current state is $h_t^x = L(h_{t-1}^x, \mathbf{M}_c^{x_t})$.
- *Temporal Attention Loss With Global LSTM Output* (*Global T-AL*). We average the LSTM output over videos to estimate temporal attention loss. The global temporal attention loss is formulated as $\mathcal{L}_{\mathcal{T}-\mathcal{AL}}(\mathrm{X},\mathrm{Y}) = \left\| \frac{1}{T}\sum_{t=1}^{T}\left(h_t^x - h_t^y\right)\right\|_2$.

In addition to the above baselines, we also compare our method with the state-of-the-art method ActorObserverNet (AONet) [4]. AONet learns a shared representation between first- and third-person videos with a Siamese-like network. We re-train their network with default parameters based on our filtered dataset.

5.3 Evaluation of the Joint Attention Network (JANet)

Quantitative Analysis. The quantitative results of different methods on the two tasks of pairs discrimination and best-match moment localization are given in Table 1. It can be seen that the JANet significantly outperforms AONet [4] on both two tasks. The performance improvement probably owes much to the predicted joint attention regions in shared representation learning. While AONet [4] attempted to learn the shared representation directly from CNN-based features, our method learns the shared representation in a more efficient and reliable way by exploiting the joint attention regions which capture the same action recorded from two viewpoints.

The ablation study results are also shown in the lower part of Table 1. We find that the removal of attention information from either side of first- or third-person videos in triplet loss leads to obvious performance drop (close to [4]), demonstrating the effectiveness of attention information in shared representation learning. Moreover, the removal of attention information from third-person video in self-supervised attention learning leads to more serious performance drop than similar removal from first-person video, which indicates that the attention information of third-person video plays a more decisive role. Most importantly, the performance degrades dramatically for *Without SA* when attention is learned independently from both viewpoints, indicating the critical role of the self-supervised joint attention learning in our full model. Overall, the ablation study results show that in shared representation learning not only the attention information is needed but also the attention from different viewpoints should be learned jointly.

TABLE 1
Quantitative Results of JANet for Algorithmic Evaluation

Method	Classification accuracy↑	Alignment error↓
AONet [4]	51.8	6.5
Without SA	52.1	6.8
CNN SA /1	88.4	5.2
CNN SA /3	61.1	5.0
CNN TL /1	52.4	6.3
CNN TL /3	56.9	6.2
Low-level TW	85.7	7.1
JANet	**90.6**	**4.5**

The performance of pairs discrimination task is measured by classification accuracy (in %). The performance of best-match moment localization task is measured by alignment error (in seconds) between best-match moment and ground-truth moment.

Qualitative Analysis. Qualitative results are shown in Fig. 8. We visualize the ROAs predicted by AONet [4], one of our baselines (Without SA), and our JANet architecture. For AONet, we visualize activations of the last convolutional layer to show which regions the network focuses on. For our baseline and JANet, ROAs are visualized by the weighted average of feature maps based on the generated channel attention vectors. It can be seen that [4] tends to focus on either the image center or the visually salient object. Taking the first group (column 1 - 2 of Fig. 8) for example, while [4] focuses on the center region of the first-person image and human body of the third-person image, our method successfully locates the shared ROAs in both images around the object of a saucepan.

As for the baseline of *Without SA* which learns attention independently for first- and third-person videos, its learned attention becomes unreliable and fails to predict the ROAs shared between two viewpoints. For example, the second group (column 3 - 4 of Fig. 8) demonstrates that the person is playing a mobile game, and the shared ROAs are around the region of mobile phone and hands. However, the ROA of the third-person image predicted by *Without SA* is located on the person's leg unrelated to the performed action. Overall, these results demonstrate that the joint attention learning via self-supervised is essential for shared representation learning between first- and third-person videos.

5.4 Evaluation of the Temporal Joint Attention Network (T-JANet)

Quantitative Analysis. Table 2 shows that quantitative results of T-JANet and its baselines. Compared with JANet, T-JANet learns temporal transition of joint attention from

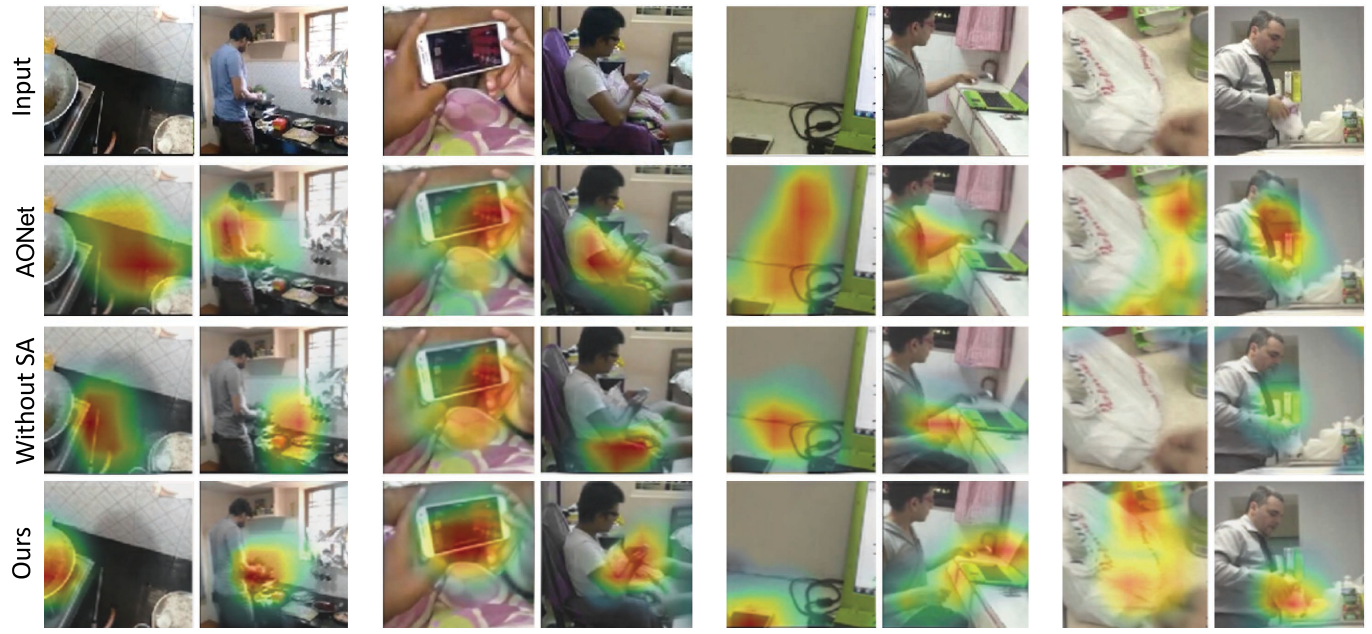


Fig. 8. Visualization of the predicted ROAs from AONet [4], our baseline of *Without SA* and our JANet. We demonstrate four groups of image pairs, each containing frames from the first- and third-person viewpoints respectively.

TABLE 2
Quantitative Results of T-JANet for Algorithmic Evaluation

Method	Classification accuracy↑	Alignment error↓
TTL Without Avg	92.3	2.1
Without T-AL	87.8	3.3
T-AL Without Avg	91.6	3.3
Global T-AL	90.3	5.0
JANet	90.6	4.5
T-JANet	**95.4**	**1.6**

The performance of pairs discrimination task is measured by classification accuracy (in %). The performance of best-match moment localization task is measured by alignment error (in seconds) between best-match moment and ground-truth moment.

cross-view videos and significantly improves the performance with highest classification accuracy (95.4) percent and lowest alignment error (1.6 s) on the two tasks of pairs discrimination and best-match moment localization respectively. This demonstrates that both spatial and temporal information are needed for shared representation learning between two viewpoints.

The influence of temporal constraint is analyzed in ablation study. *Without T-AL* removes the temporal attention loss and leads to a performance drop, indicating the importance of temporal constraint. *TTL Without Avg* and *T-AL Without Avg* keep the same LSTM network but only use information of current frame as input. The results of these two baselines show a slight performance drop compared with that of T-JANet, indicating that temporal information from adjacent frames is also important.

Qualitative Analysis. Qualitative comparison of JANet and T-JANet is presented in Fig. 9. We visualized the predicted ROAs of JANet and T-JANet in three groups. Since T-JANet exploits the high-level attention transition temporally, it can predict shared ROAs more consistently than JANet under complex situations. The first group (column 1 - 3 of Fig. 9), as an example, illustrates that the shared ROAs are always located on region of sandwich. Due to independent process of each input frame, JANet wrongly predicts the ROA of the first-person view on the woman's leg in the third frame (highlighted in the red dot of Fig. 9a). On the contrary, with the help of temporal information, T-JANet successfully predicts the shared ROAs across different frames. For example, the third group shows that the ROA of the second first-person frame (column 8 of Fig. 9) is occasionally out of view due to instant head motion. In this case, our T-JANet can also correctly predict the ROA with the information of previous frames.

Moreover, we also conduct attention estimation on other unseen datasets (ActivityNet [49] and 1st-3rd dataset [8]) to examine the generalizability of our attention model. ActivityNet dataset is a large-scale video benchmark that covers a wide range of complex human activities from different viewpoints. It consists of a total of 849 video hours which are recorded by different people in different scenes. The 1st-3rd dataset is composed of sets of three synchronized videos (two first-person videos and one third-person video) ranging between 5-10 minutes for the task of person-video identification. The attention estimation results on the two datasets are shown in Fig. 10. On the ActivityNet dataset, since our method exploits the attention transition temporally, it can predict shared ROAs consistently under various situations. On the 1st-3rd dataset (Fig. 10a), the attention estimation in the third-person video (top row) demonstrates more stable results than in the first-person video (bottom row). The reason might be that the first-person videos in the 1st-3rd dataset are recorded by Xiaoyi Yi Action Cameras[1], which are chest-mounted and provide different views from the head-mounted cameras used in the Charades-Ego dataset.

1. http://www.xiaoyi.com/en/specsen.html

Fig. 9. Visualization of predicted ROAs from our JANet (a) and T-JANet (b). We presented video actions with predicted ROAs in both first- and third-person viewpoints (from top to bottom: third-person perspective and corresponding first-person perspective). ROAs are visualized with heatmaps on input images. The color ranges from blue to red, showing low to high attention.

Overall, these results demonstrate that temporal attention constraint plays an important role in learning joint attention and shared representation of cross-view videos.

5.5 Evaluation of Representative Applications

As shown in previous sections, our method is also capable of predicting ROAs and learn shared representation across different views. It has been studied in previous works that the estimation of attention could be used to focus on important visual regions and is important for various computer vision tasks such as object segmentation and action recognition. Our method is able to predict ROAs without explicit supervision. Therefore, it can be used as a fundamental tool and applied to tackle a variety of computer vision tasks. In this section, we conduct experiments on four representative applications to validate the efficiency of our method. The representative applications are defined as follows:

- *Gaze Prediction (Single-View Input / JANet Framework).* Third-person video frames are taken as input. Gaze direction is predicted by combining the predicted ROA from JANet and the head position inferred with a 2D body pose estimation algorithm.
- *Image Co-Segmentation (Cross-View Input / JANet Framework).* Images of two views are taken as input. The predicted ROAs from JANet are combined with a superpixel algorithm to locate the object region of joint attention between two views.
- *Video Summarization (Cross-View Input / T-JANet Framework).* Videos of two views are taken as input. Key frames are selected on account of the per-frame importance score which is computed based upon similarity of features extracted from T-JANet.
- *Zero-Shot Action Recognition (Cross-View Input / T-JANet Framework).* Videos of a target view are taken as input. Features are extracted with T-JANet and actions are recognized only based on training data (videos with action labels) from a different view.

Note that in the case of cross-view image co-segmentation, there is no cross-view data exchange during calculation. For the case of video summarization and zero-shot action recognition, our method takes into account both first- and third-person viewpoints. The calculation of each views is interdependent.

(a) Videos from 1st-3rd dataset [49] (b) Videos from Activity Net

Fig. 10. Visualization of the predicted ROAs with T-JANet on unseen 1st-3rd dataset [8] (a) and ActivityNet dataset [49] (b). The top row shows the third-person view and the bottom row shows the first-person view.

5.5.1 Gaze Prediction

Predicting where a person looks (gaze) in a third-person video is important for video analysis and information retrieval. Existing single-view based gaze prediction methods [50], [53], [54], [55], [56], [57] locate the human eyes first, and then estimate the gaze direction based on the pupil and head's direction. Different from previous methods, we predict the gaze as follows: first, we locate the position of

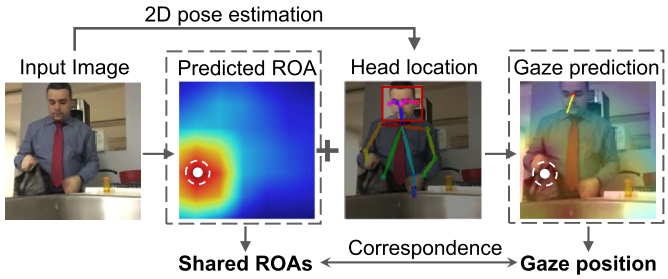


Fig. 11. Illustration of our gaze prediction framework. The predicted ROA of third-person frame is used to locate the gaze position of the actor. The white dot indicates the gaze direction inferred from predicted ROA.

human head based on the pose estimation method [54]. Next, we use the third-person frame's ROA, which is inferred from our model, as the place where a person looks. The pipeline of our gaze prediction approach is demonstrated in Fig. 11.

We compare our method with the state-of-the-art single-view based gaze prediction method AC17 [50] with its default parameters. Since Charades-Ego dataset does not provide ground-truth gaze annotations, we consider two ways to evaluate gaze prediction. 1) We manually annotate a subset of third-person video frames in the Charades-Ego dataset to generate a new dataset for evaluating gaze prediction (we call it "3rdGaze" dataset). 3rdGaze dataset has a total of 1059 annotated images which cover 19 scenes and 182 objects categories. All images in the 3rdGaze dataset are annotated with 2D eye positions and gaze points. We use average angular error (AAE) as our evaluation metric, which is formulated as: $\frac{1}{n}\sum_i^n \|\alpha_i - \beta_i\|$, where α_i is the estimated gaze angle and β_i is the annotated gaze angle of i-th image. n is the number of images in the 3rdGaze dataset. 2) We also conduct user study to provide a subjective evaluation of different methods. A corpus of 100 participants (42 females and 58 males) with diverse backgrounds was recruited to participate in the user study. Fifty third-person video frames randomly selected from the Charades-Ego datasets are used to form a questionnaire. Participants are asked to examine the gaze prediction results of two methods and choose their preferred ones. User preference ratio is used as the evaluation metric.

The left part of Table 3 shows the gaze prediction results with two evaluation metrics. It can be seen that 89.3 percent of participants prefer our method compared with [50]. The average angular error of our method is much lower than AC17 [50], indicating the effectiveness of our method for gaze prediction. Fig. 12 illustrates the qualitative comparison

TABLE 3
Quantitative Results of Gaze Prediction and Image Co-Segmentation

Gaze prediction			Image co-segmentation	
Method	Preference↑	AAE↓	Method	Preference↑
AC17 [50]	10.7%	30.6°	CH18 [52]	19.3%
Ours	**89.3%**	**16.7°**	**Ours**	**80.6%**

User preference ratio (Preference, higher is better) is used as the evaluation metric for both tasks, and average angular error (AAE, lower is better) is used as the evaluation metric for gaze prediction.

Fig. 12. Examples of gaze prediction results. The top row indicates the gaze prediction results of AC17 [50]. The bottom row indicates the gaze prediction results of our approach. N/A denotes AC17 [50] failing to predict the direction of human gaze.

between our method and [50]. With predicted ROAs, our method achieves obviously better results than [50] which relies on face landmark detection. Furthermore, our method could robustly predict a gaze position when the human face cannot be detected in an input image.

5.5.2 Image Co-Segmentation

Here we show that our method could be easily extended to solve image co-segmentation task in an unsupervised manner. Different from previous methods [52], [58], [59], the key of our approach is that we utilize the predicted ROAs of both first- and third-person video frames to guide co-segmentation. First, candidate segments are extracted by an unsupervised image segmentation method (here we adopt SLICO [51]). Then, we choose two segments from candidate segments of two images as output which are near the center of ROAs and are also visually similar to each other. The pipeline of our co-segmentation framework is demonstrated in Fig. 13.

We compare our method with the state-of-the-art image co-segmentation method of CH18 [52]. Similar to Section 5.5.1, we conduct a user study for quantitative evaluation. We randomly sampled 25 pairs of first- and third-person video frames to compare the co-segmentation results of [52] and those of our method.

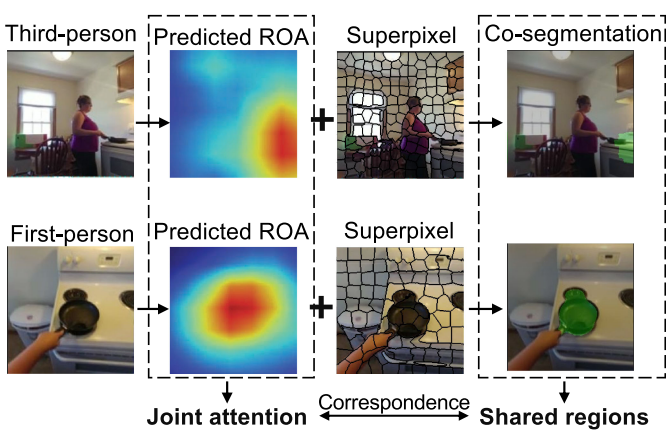


Fig. 13. Illustration of image co-segmentation framework. We adopt superpixel algorithm [51] to group pixels into perceptually meaningful atomic regions. The joint attention is used to indicates the shared regions between input pairs.

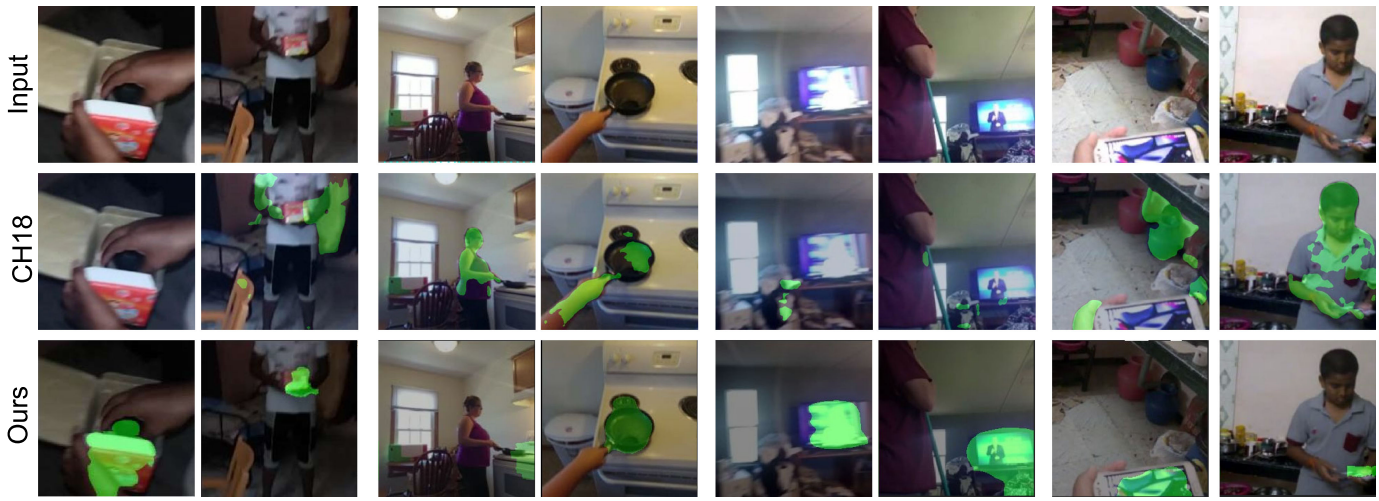


Fig. 14. Examples of co-segmentation results. The first row depicts input image pairs from two perspectives, the second row displays the object co-segmentation results of CH18 [52], and the third row displays the results of our approach.

The results of user preference ratio are shown in the right part of Table 3. It can be seen that 80.6 percent of participants prefer our method compared with [52]. Quantitative comparison is shown in Fig. 14. With the guidance of attention information, our method significantly outperforms [52].

5.5.3 Video Summarization

In this part, we show that joint attention predicted from first- and third-person videos could be exploited to discover important moments that describe the potential undergoing activity. Thus, our joint attention provides an effective solution for multi-view video summarization. The state-of-the-art video summarization methods [60], [61], [62], [63], [64], [65] often take single-view video as input, estimate importance scores per frame and create a video summary consisting of a small subset of frames. In contrast, we compute per-frame scores of joint attention for paired first- and third-person videos by computing the similarity between channel attention vectors of two videos. We assign high importance scores to frames that have high scores of joint attention, based on which a subset of frames with high importance scores above a threshold is selected as key frames for summarization. The pipeline of our video summarization framework is demonstrated in Fig. 15.

We compare our multi-view video summarization method with two state-of-the-art methods (ZQ18 [60] and GG14 [64]) on Charades-Ego dataset. We also conducted performance comparison based on per-frame importance scores generated from AONet [4] and our JANet. Note that for AONet, we compute importance scores by evaluating the similarity between basic CNNs feature maps. We use the annotated action clips as ground-truth of video summary and use average precision and F1 score as the evaluation metric. Since action clips that are too short or too long are inappropriate for task evaluation, we removed those clips in two ways. 1) We consider an action clip that lasts less than one second as an accidental action, which is not related to the whole activity and thus is removed. 2) We also remove videos in which more than 95 percent of the frames belong to a single action, as summarization is not needed in these videos.

Quantitative results are shown in Table 4. Since GG14 [64] predicts frame importance scores based on specific image features such as facial landmark detectors and motion, it shows the worst result among the baselines. In contrast, our T-JANet based method could reliably detect more important moments by joint attention information with the help from additional third-person viewpoint. We show qualitative example of our T-JANet result in Fig. 16.

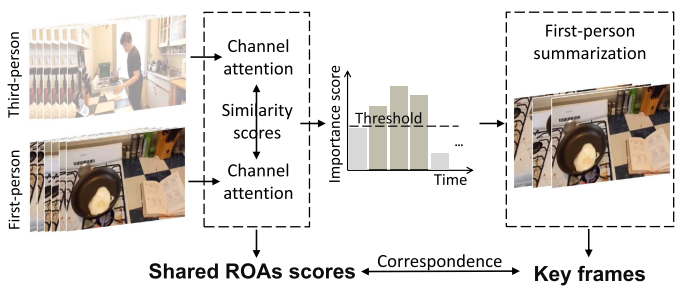


Fig. 15. Illustration of our video summary framework. The shared ROAs are used to estimate the importance scores of each frames by computing the similarity between channel attention vectors of two videos. Then, key frames are selected based on per-frame importance scores.

5.5.4 Zero-Shot Action Recognition

In this section, we present results of cross-view action recognition in a zero-shot setting in order to show how the knowledge from a source viewpoint can be transferred to a target viewpoint with shared representation learned between the two views.

TABLE 4
Quantitative Comparison With State-of-the-Art Video Summarization Methods on Charades-Ego Dataset

Method	Average precision	F1 score
ZQ18 [60]	57.7	68.0
GG14 [64]	52.3	58.1
AONet [4]	61.3	64.8
Our JANet	61.3	66.3
Our T-JANet	**64.4**	**69.8**

Average precision and F1 score (in %) are used as the evaluation metrics.

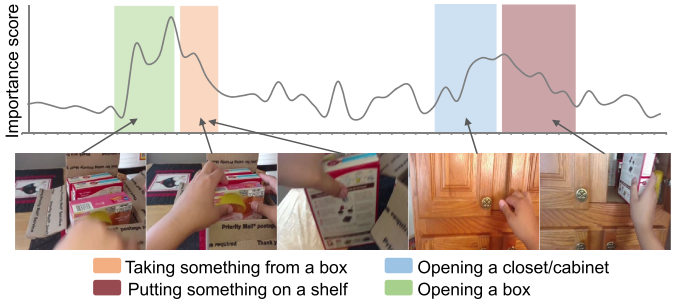


Fig. 16. Examples of video summaries generated by our T-JANet on Charades-Ego dataset. The gray line shows the per-frame importance scores. Color-intensified areas indicate frames of different actions. The details of action labels are shown on the bottom.

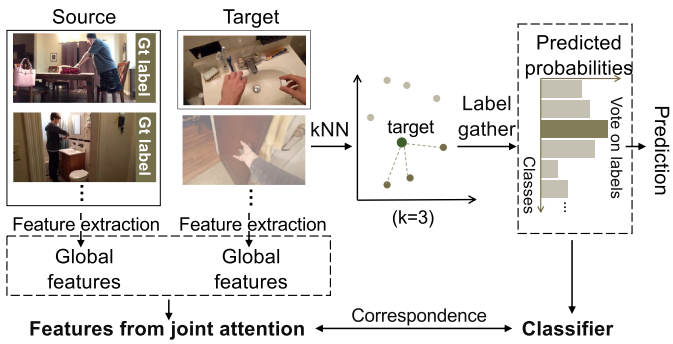


Fig. 17. Illustration of our zero-shot action recognition framework. The extracted features are used to predict the classification probability of a target sample with kNN.

Zero-shot action recognition is implemented with a k-nearest neighbors algorithm (k-NN) [66]. Assuming that the source domain data is made up of third-person video clips (one-second) with ground-truth action labels, and the target domain data is comprised of first-person video clips without labels. Given a first-person video clip, k nearest third-person video clips are searched out based on the L2 distance between the feature vectors of video clips. Action category with most votes from the k nearest third-person video clips is then assigned as final prediction. Following the evaluation setup from Charades [67], multi-class mean average precision (mAP) is used as the evaluation metric. The pipeline of this application is shown in Fig. 17.

Performance comparison is conducted based on features extracted from three network architectures: AONet [4], JANet and T-JANet. In addition, Sigurdsson *et al.* [4], [68] implemented two baselines: one is trained on both first- and third-person videos and tested on either first- or third-person videos using the models from [69]. We denoted this baseline as "SG18". The other one implemented zero-shot first-person action recognition by adding a classification loss and then jointly training the network. This baseline is denoted as "AONet-cls". Quantitative results are shown in Table 5. AONet outperforms AONet-cls by a large margin, which indicates that non-parametric feature matching is more suitable for cross-view action recognition than training an additional action classifier. The proposed T-JANet achieves best performance among the three architectures, validating the advantage of learning with spatial-temporal joint attention. Fig. 18 demonstrates the results with different values of k. It can be seen that performance converges nearly with $k \geq 5$.

TABLE 5
Quantitative Results of Zero-Shot Action Recognition

	SG18 [68]	AONet-cls [4]	AONet	JANet	T-JANet
$3^{rd} \rightarrow 1^{st}$	19.5	25.9	43.2	52.1	**58.5**
$1^{st} \rightarrow 3^{rd}$	17.5	-	38.9	53.1	**60.2**

The performance is measured by video-level mAP (in %). $3^{rd} \rightarrow 1^{st}$ indicates that we use third-person videos as training samples, and the first-person videos are used to test. $1^{st} \rightarrow 3^{rd}$ indicates that we use first-person videos as training samples, and the third-person videos are used to test.

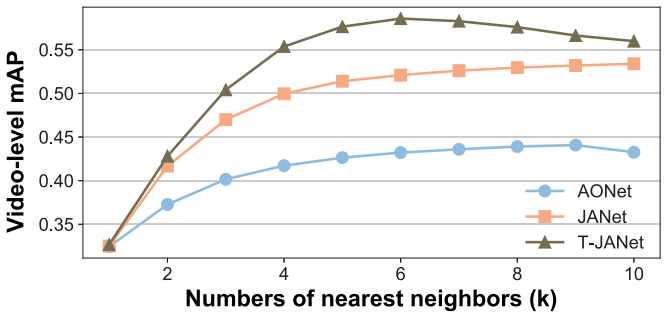


Fig. 18. Zero-shot action recognition performance with different k values ($k \in [1, 10]$) based on AONet [4], JANet and T-JANet.

6 Conclusion and Future Work

In this paper, we propose a method to effectively learn a shared representation for co-analysis of the first- and third-person videos. Our key idea is to learn joint attention for linking these two viewpoints, with the assumption that shared representation should correspond to the joint attention regions. A novel representation learning framework with a self-supervised attention learning module is developed to learn joint attention spatially and temporally. Experiment results on a public dataset show that our proposed method significantly outperforms the state-of-the-art method on two cross-view video matching tasks. Additional experiments are conducted to demonstrate the benefits of our work for various applications.

In the future, we will deploy our method to more scenes. In addition, based on the shared representation learning between first- and third-person videos, our research interests also include synthesizing first-person video in accordance with the input of a third-person video.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61972012, Grant 61732016 and Grant 61906064. Huangyue Yu and Minjie Cai contributed equally to this work.

References

[1] M. Cai, F. Lu, and Y. Gao, "Desktop action recognition from first-person point-of-view," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1616–1628, May 2018.
[2] Y. Huang, M. Cai, Z. Li, F. Lu, and Y. Sato, "Mutual context network for jointly estimating egocentric gaze and action," *IEEE Trans. Image Process.*, vol. 29, pp. 7795–7806, Jul. 2020.
[3] M. Cai, F. Lu, and Y. Sato, "Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14380–14389.

[4] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and observer: Joint modeling of first and third-person videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7196–7404.

[5] H. Yu, M. Cai, Y. Liu, and F. Lu, "What I see is what you see: Joint attention learning for first and third person video co-analysis," in *Proc. Int. Conf. Multimedia*, 2019, pp. 7396–7404.

[6] R. Yonetani, K. M. Kitani, and Y. Sato, "Ego-surfing first-person videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5445–5454.

[7] S. Ardeshir and A. Borji, "Ego2Top: Matching viewers in egocentric and top-view videos," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 1–16.

[8] C. Fan *et al.*, "Identifying first-person camera wearers in third-person videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4734–4742.

[9] M. Xu, C. Fan, Y. Wang, M. S. Ryoo, and D. J. Crandall, "Joint person segmentation and identification in synchronized first-and third-person videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 656–672.

[10] H. I. Ho, W. C. Chiu, and Y. C. F. Wang, "Summarizing first-person videos from third persons' points of views," in *Proc. Eur. Conf. Comput. Vis.*, 2017, pp. 1–10.

[11] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 639–655.

[12] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2235–2244.

[13] D. Damen *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–19.

[14] A. R. Lejbolle, B. Krogh, K. Nasrollahi, and T. B. Moeslund, "Attention in multimodal neural networks for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 292–2928.

[15] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 369–378.

[16] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2285–2294.

[17] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.

[18] T. Chen *et al.*, ""Factual" or "emotional": Stylized image captioning with adaptive learning and attention," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 527–543.

[19] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 68–86.

[20] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[21] W. Liu, J. Chen, C. Li, C. Qian, X. Chu, and X. Hu, "A cascaded inception of inception network with attention modulated feature fusion for human pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7170–7177.

[22] E. Parisotto, D. Singh Chaplot, J. Zhang, and R. Salakhutdinov, "Global pose estimation with an attention-based recurrent network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 350–359.

[23] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1831–1840.

[24] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *TIP*, vol. 27, no. 3, pp. 1487–1500, 2018.

[25] Y. Li and Y. Wang, "A multi-label image classification algorithm based on attention model," in *ICIS*, pp. 728–731, 2018.

[26] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov.1998.

[27] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[28] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–17.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 3–19.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[31] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[32] S. Hu, M. Feng, R. M. Nguyen, and G. H. Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7258–7267.

[33] N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 1–7.

[34] K. Regmi and A. Borji, "Cross-view image synthesis using conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1–10.

[35] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng, "Multi-view image generation from a single-view," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 383–391.

[36] J. L. Elman, "Finding structure in time," *Cognitive Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1725–1780, 1997.

[38] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[39] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Int. Conf Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[40] O. Costilla-Reyes, R. Vera-Rodríguez, P. Scully, and K. B. Ozanyan, "Analysis of spatio-temporal representations for robust footstep recognition with deep residual neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 285–296, Feb. 2019.

[41] H. Xu, A. Das, and K. Saenko, "Two-stream region convolutional 3D network for temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2319–2332, Oct. 2019.

[42] L. Wang *et al.*, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[43] Y. Wang, H. Huang, C. Wang, T. He, J. Wang, and M. Hoai, "Gif2video: Color dequantization and temporal interpolation of GIF images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1419–1428.

[44] A. Furnari and G. M. Farinella, "Egocentric action anticipation by disentangling encoding and inference," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3357–3361.

[45] E. Ng, K. Grauman, and D. Xiang, "You2Me : Inferring body pose in egocentric video via first and second person interactions by evonne," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9890–9900.

[46] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9029–9038.

[47] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–4

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

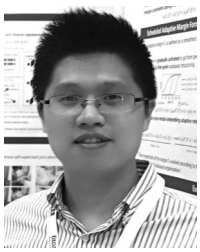
[49] F. C. Heilbron, V. Escorcia, B. Ghanem and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.

[50] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba, "Following gaze in video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1444–1452.

[51] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

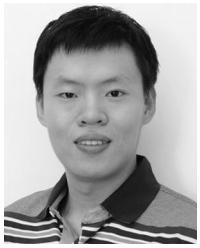
[52] H. Chen, Y. Huang, and H. Nakayama, "Semantic aware attention based deep object co-segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp 435–450.
[53] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," in *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, Mar. 2020.
[54] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.
[55] F. Lu, X. Chen, and Y. Sato, "Appearance-based gaze estimation via uncalibrated gaze pattern recovery," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1543–1553, Apr. 2017.
[56] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2033–2046, Oct. 2014.
[57] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI*, 2020, pp. 10623–10630.
[58] R. Quan, J. Han, D. Zhang, and F. Nie, "Object co-segmentation via graph optimized-flexible manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 687–695.
[59] P. Mukherjee, B. Lall, and S. Lattupally, "Object cosegmentation using deep siamese network," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell.*, 2018, pp. 1–5.
[60] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7582–7589.
[61] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool, "Query-adaptive video summarization via quality-aware relevance estimation," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 582–590.
[62] H. Jin, Y. Song, and K. Yatani, "Elasticplay: Interactive video summarization with dynamic time budgets," in *Proc.25th ACM Int. Conf. Multimedia*, 2017, pp. 1164–1172.
[63] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc.25th ACM Int. Conf. Multimedia*, 2017, pp. 863–871.
[64] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp 505–520.
[65] K. Zhang, W. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016. pp. 1–24.
[66] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.
[67] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *Proc. Eur. Conf. Comput. Vis.*, 2016, pp 510–526.
[68] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," 2018, *arXiv:1804.09626*. [Online]. Available: http://arxiv.org/abs/1804.09626
[69] G. A. Sigurdsson, S. K. Divvala, A. Farhadi, and A. Gupta, "Asynchronous temporal fields for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 585–594 .

Huangyue Yu received the BS degree in digital media technology, in 2018. She is currently working toward the MS degree from the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University. Her research interests include computer vision, human-computer interaction, and video analysis.



Minjie Cai received the BS and MS degrees in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011, respectively and the PhD degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2016. He is currently an assistant professor at the College of Computer Science and Electronic Engineering, Hunan University. His research interests include computer vision, multimedia, and human-computer interaction.

Yunfei Liu is currently working toward the PhD degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University. His research interests include computer vision, computational photography, and image processing.



Feng Lu (Member, IEEE) received the BS and MS degrees in automation from Tsinghua University, in 2007 and 2010, respectively, and the PhD degree in information science and technology from the University of Tokyo, in 2013. He is currently a professor at the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision, human-computer interaction, and augmented intelligence.