Synchronization is All You Need: Exocentric-to-Egocentric Transfer for Temporal Action Segmentation with Unlabeled Synchronized Video Pairs

Camillo Quattrocchi[1], Antonino Furnari[1,2], Daniele Di Mauro[2],
Mario Valerio Giuffrida[3], and Giovanni Maria Farinella[1,2]

[1] Department of Mathematics and Computer Science, University of Catania, Italy
[2] Next Vision s.r.l., Italy
[3] School of Computer Science, University of Nottingham, United Kingdom

Abstract. We consider the problem of transferring a temporal action segmentation system initially designed for exocentric (fixed) cameras to an egocentric scenario, where wearable cameras capture video data. The conventional supervised approach requires the collection and labeling of a new set of egocentric videos to adapt the model, which is costly and time-consuming. Instead, we propose a novel methodology which performs the adaptation leveraging existing labeled exocentric videos and a new set of unlabeled, synchronized exocentric-egocentric video pairs, for which temporal action segmentation annotations do not need to be collected. We implement the proposed methodology with an approach based on knowledge distillation, which we investigate both at the feature and Temporal Action Segmentation model level. Experiments on Assembly101 and EgoExo4D demonstrate the effectiveness of the proposed method against classic unsupervised domain adaptation and temporal alignment approaches. Without bells and whistles, our best model performs on par with supervised approaches trained on labeled egocentric data, without ever seeing a single egocentric label, achieving a +15.99 improvement in the edit score (28.59 vs 12.60) on the Assembly101 dataset compared to a baseline model trained solely on exocentric data. In similar settings, our method also improves edit score by +3.32 on the challenging EgoExo4D benchmark. Code is available here: https://github.com/fpv-iplab/synchronization-is-all-you-need.

Keywords: Temporal Action Segmentation · Egocentric Vision · View Adaptation

1 Introduction

Temporal Action Segmentation (TAS) consists in dividing a video into meaningful temporal segments or intervals, where each segment corresponds to a distinct action or activity. TAS is important for different applications, ranging from collaborative robotics to the analysis of activities of daily living [24]. TAS has
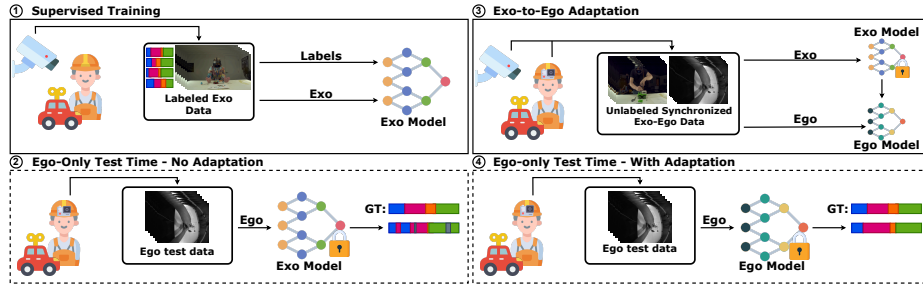
Fig. 1: We consider the problem of transferring a temporal action segmentation model from an exocentric to an egocentric setup. ① We assume that a set of labeled exocentric videos is available to train an exocentric temporal action segmentation model. ② When tested on egocentric data, the exocentric model exhibits poor performance due to domain shift. ③ We propose to use an exocentric-to-egocentric adaptation process to adapt the exocentric model using a set of synchronized unlabeled exocentric-egocentric video pairs. ④ At test time, the model is able to operate on egocentric data, with no access to exocentric videos.

been often studied in the context of *exocentric* (or third-person) vision, where videos are collected through fixed cameras [23, 38, 43, 50]. A line of work also considers the *egocentric* (or first-person) domain, in which videos are collected by wearable devices [36, 39, 40, 44]. Providing a privileged view on human-object interactions and object manipulation activities, the application of TAS to egocentric videos is relevant in different contexts ranging from work scenarios and assembly lines [4, 39] to understanding daily activities [9, 28]. Indeed, the exploitation of TAS in such contexts can be useful to analyze the user's behavior and provide assistive services through wearable devices.

Despite these potential benefits, current systems are mostly designed and tuned on exocentric videos collected and labeled by expert annotators. Due to the domain gap between exocentric and egocentric views, these systems tend to perform poorly when tested on videos collected through wearable cameras. TAS algorithms can be adapted to the new viewpoint through a supervised approach which consists in training or fine-tuning models on new egocentric videos of subjects performing the target activities, which have been purposely collected and labeled, a time-intensive and expensive process. An alternative approach would consist in using unsupervised domain adaptation algorithms [49] to train the model using pre-existing labeled exocentric videos and newly collected unlabeled egocentric videos. This approach is more convenient than the above-discussed standard supervised scheme as it only relies on unlabeled egocentric videos, thus avoiding the expensive annotation process when the model is adapted to the egocentric scenario. However, as we also show in this paper, models trained with unsupervised domain adaptation tend to be affected by a residual domain shift that prevents full generalizability to the target domain (i.e., egocentric videos).

We observe that, in real application scenarios, it is inexpensive to collect an additional set of synchronized unlabeled *exocentric-egocentric video pairs* of sub-

jects performing the activities of interest. For example, in a factory, workers can be equipped with wearable devices recording their activities while the already installed fixed cameras acquire synchronized videos of the same activities.[4] The collected video pairs are naturally synchronized and they are not manually annotated for the temporal action segmentation task. Therefore, we investigate whether such unlabelled synchronized video pairs can provide a suitable form of supervision to adapt an existing exocentric TAS model to the egocentric domain (see Figure 1). Based on the synchronized adaptation set, algorithms will aim to leverage domain-specific information learned from the source exocentric dataset to generate a model that performs well in the target egocentric video sequences. To tackle the proposed task, we introduce a novel exocentric-to-egocentric adaptation methodology based on knowledge distillation, investigating its application at the level of the feature extractor and downstream TAS model. We perform experiments on two different view adaptation settings of the Assembly101 [39], which comprises multiple synchronized egocentric and exocentric video sequences of subjects assembling toys, and on the recently released EgoExo4D dataset [16], which includes multi-view synchronized sequences of subjects performing diverse activities. Experiments confirm that synchronized video pairs provide a valuable form of supervision, with our best model bringing improvements of +15.99 and +11.64 edit score on Assembly101 in the two view transfer settings, compared to baselines trained solely on exocentric data, and recovering the exo-ego domain gap in EgoExo4D, with improvements of up to +3.32 in edit score.

Our contributions are as follows: 1) we propose and investigate a new exocentric-to-egocentric adaptation task which aims to transfer an already trained exocentric TAS system to egocentric cameras in the presence of cross-view synchronized video pairs; 2) we design and benchmark an adaptation methodology which performs knowledge distillation at the level of the feature extractor and downstream TAS model; 3) we benchmark the proposed task on the Assembly101 [39] and EgoExo4D [16] datasets, reporting results on different view adaptation settings. Experiments confirm the effectiveness of the proposed methodology with significant improvements over current approaches, suggesting that synchronization allows to recover the performance of supervised methods, without relying on egocentric labels. Code is available here: `https://github.com/fpv-iplab/synchronization-is-all-you-need`.

2   Related Work

Temporal Action Segmentation The literature on Temporal Action Segmentation is rich, with many approaches based on classic [44] or modern [30] techniques. Among recent representative methods, the authors of [24] first proposed an encoder-decoder model based on temporal convolutional networks. The authors of [26] later proposed a multistage architecture based on temporal convolutional networks that refines the predicted actions in multiple stages. In [23],

---

[4] As we show in our experiments, our method works also when videos are not perfectly synchronized, hence sophisticated synchronization systems are not needed.

it was presented an approach based on a recurrent neural network that models combinations of discriminative representations of subactions, allowing temporal alignment and inference over long sequences. The authors of [42] proposed C2F-TCN, a temporal encoder-decoder architecture based on an implicit ensemble of multiple temporal resolutions to mitigate sequence fragmentation. The study in [43] showed that it is possible to learn higher-level representations to interpret long video sequences. The authors of [50] proposed DTL, a framework that uses temporal logic to constrain the training of action analysis models improving the performance of different architectures. The authors of [30] use denoising diffusion models to iteratively refine predicted actions starting from random noise, using video features as conditions. The reader is referred to [10] for a recent survey on the topic of Temporal Action Segmentation. We follow [39] and consider a Coarse to Fine Temporal Convolutional Network (C2F-TCN) [42] as a temporal action segmentation approach.

Exocentric-to-Egocentric Model Transfer Few previous works considered the problem of transferring knowledge from exocentric to egocentric views. The authors of [41] proposed the Charades-Ego dataset containing pairs of first- and third-person synchronized videos, labeled for action recognition. A model jointly learning to represent those two domains with a triplet loss was also proposed. The approach presented in [53] learns a shared representation of exocentric and egocentric videos by modeling joint attention maps to link the two viewpoints. The authors of [27] proposed to leverage latent signals in third-person videos that are predictive of egocentric-specific properties, such as camera position, manipulated objects, and interactions with the environment to kickstart the training of an egocentric video model on exocentric data. In [51], a method to jointly learning from unpaired egocentric and exocentric videos was proposed, with the aim to leverage temporal alignment as a self-supervised learning objective and learn view-invariant features. The work in [25] introduced a method to extrapolate affordance grounding from image-level labels on exocentric images and transfer them to egocentric views of the same objects. Previous approaches focused on exocentric-to-egocentric transfer in the absence of synchronized video pairs and considered tasks other than Temporal Action Segmentation. Ours is the first systematic investigation on unlabeled synchronized exo-ego video pairs as self-supervision for exo-to-ego model transfer for Temporal Action Segmentation.

Domain Adaptation Domain Adaptation methods aim to leverage the knowledge gained from a specific source domain to improve performance of a machine learning model on a target domain in the presence of a distribution shift [8]. In the unsupervised domain adaptation setting, models assume the presence of source labeled data and target unlabeled data during training. In recent years, many domain adaptation techniques have been proposed for several computer vision tasks, including some related to video understanding [6, 21]. Different works focused on the adaptation of action recognition models when source and target domains are given by different environments. Among these works, the authors of [2] introduced two large-scale datasets for video domain adaptation for action recognition, *UCF-HMDB$_{full}$* and *Kinetics-Gameplay*, comprising both real and

virtual domains, and proposed the Temporal Attentive Adversarial Adaptation Network (TA3N) that simultaneously attends, aligns, and learns temporal dynamics across domains, achieving state-of-the-art performance. In [33], the RGB and optical flow modalities are leveraged to reduce the domain shift arising from training and testing action recognition models in different environments. The authors of [22] investigated video domain adaptation for action recognition in a cross-modal contrastive learning framework. The authors of [32] introduced *ADL-7* and *GTEA-KITCHEN-6*, two egocentric domain-adaptation datasets for action recognition, and proposed a Channel Temporal Attention Network (CTAN) architecture. The work of [5] demonstrated promising performance on the task of adapting first- and third-person videos for drone action recognition. The problem of domain adaptation for temporal action segmentation was investigated in [3], where different domains arise from different environments. Previous works mainly focused on action recognition [2, 5, 33] and considered different environments or subjects as a form of distribution shift [2, 32, 33] with few works tackling exocentric-to-egocentric adaptation as a domain adaptation problem [5] or considering temporal action segmentation as a target task [3]. Also, previous works considered unpaired inputs from the different domains, which limit the amount of supervision that can be leveraged for the model transfer. Our investigation is first to consider exo-to-ego adaptation for temporal action segmentation in the presence of synchronized unlabeled video pairs, which, as we show in the experiments, provide a strong form of self-supervision for adaptation.

Knowledge Distillation Early knowledge distillation methods were designed to transfer knowledge from a resource intensive neural network (the "teacher") to a less computationally demanding network (the "student") [19]. These techniques exploited the teacher model's logits to supervise the student network through a Kullback–Leibler divergence loss on the probability distributions predicted by the two models. An alternate set of strategies proposed to align the intermediary activations of teacher and student networks to provide a more accurate distillation signal [20, 37]. These approaches stem from the observation that neural networks internally develop hierarchical representations of their inputs, which characterize the way final predictions are made. Another line of approaches proposed to model the relationships between activations in consecutive layers of the teacher network as a way of guiding the student's learning [35, 52]. Beyond adapting a larger model to a smaller one, knowledge distillation has also been used to improve the results of action recognition [7, 31, 45], early action prediction [48], and action anticipation [1, 12, 13, 47]. Unlike the aforementioned works, we use knowledge distillation to adapt a model trained solely on exocentric videos to generalize over egocentric videos based on a set of unlabeled synchronized exocentric-egocentric video pairs. We investigate distillation at the levels of feature extraction and TAS model, showing that unlabeled synchronized exo-ego pairs provide a strong learning objective for exo-to-ego model transfer.

3   Methodology

Problem Definition Let $\mathcal{D}_{train}^{exo} = \{(V_i^{exo}, S_i^{exo})\}_{i=1}^n$ be a collection of n labeled exocentric videos, where V_i^{exo} is the ith exocentric video of the collection and S_i^{exo} is the corresponding manually labeled segmentation. Let $\mathcal{D}_{adapt}^{pair} = \{(V_j^{exo}, V_j^{ego})\}_{j=1}^m$ be a collection of m time-synchronized unlabeled exocentric-egocentric video pairs, designed to perform the adaptation process. Note that these video pairs can be collected with existing egocentric and exocentric systems and do not need to be labeled. We define exocentric-to-egocentric domain adaptation as the problem of training a TAS algorithm Ψ on both $\mathcal{D}_{train}^{exo}$ and $\mathcal{D}_{adapt}^{pair}$ to generalize over a test set $\mathcal{D}_{test}^{ego} = \{(V_h^{ego}, S_h^{ego})\}_{h=1}^q$ of egocentric videos V_h^{ego} and the corresponding ground truth temporal segmentations S_h^{ego} (used only for evaluation purposes). The proposed task is related to two other problems previously investigated in the literature: 1) standard supervised TAS, that directly trains on a labeled dataset of egocentric videos $\mathcal{D}_{train}^{ego}$ [10]; 2) cross-view domain adaptation [27], which trains on a labeled set of exocentric videos $\mathcal{D}_{train}^{exo}$ and an unlabeled set of egocentric videos $\mathcal{D}_{adapt}^{ego} = \{V_j^{ego}\}_{j=1}^m$ for adaptation. Differently from 1), we aim to generalize to egocentric videos using only labeled exocentric videos of $\mathcal{D}_{train}^{exo}$ and a set $\mathcal{D}_{adapt}^{pair}$ of unlabeled exocentric-egocentric synchronized video pair, whereas, differently from 2), we make use of a set of exocentric-egocentric video pairs $\mathcal{D}_{adapt}^{pair}$, which are deemed to provide a stronger supervision than the set of egocentric videos $\mathcal{D}_{adapt}^{ego}$.

3.1   Proposed Adaptation Approach

We propose a methodology to adapt an exocentric TAS model to egocentric videos that relies on the distillation of knowledge at two different levels: the feature extractor and the downstream TAS model. See Figure 2.

Algorithm Factorization Our model Ψ is factorized into $\Psi = \gamma \circ \phi$, where ϕ is a feature extractor and γ is the TAS model making predictions based on the features extracted by ϕ.

Teacher Model We first train a teacher model by optimizing the feature extractor ϕ^T and the downstream TAS algorithm γ^T on the exocentric supervised training set $\mathcal{D}_{train}^{exo}$. We train the C2F-TCN model γ^T on top of features extracted by the feature extractor ϕ^T. After training, the parameters of the teacher model are fixed. The teacher is no longer optimized during the distillation process.

Student Model The student model shares the same architecture as the teacher model, adopting the same feature extractor ϕ^S and prediction algorithm γ^S.

Distillation Process During the distillation process, the student is optimized to mimic the outputs of the teacher component. This process is carried out using the adaptation set $\mathcal{D}_{adapt}^{pair}$. No exocentric or egocentric labels are used at this time, with the only supervision being given by the exocentric videos passed through the fixed teacher. We consider two adaptation steps that can be either included separately or combined: feature distillation and TAS model distillation.
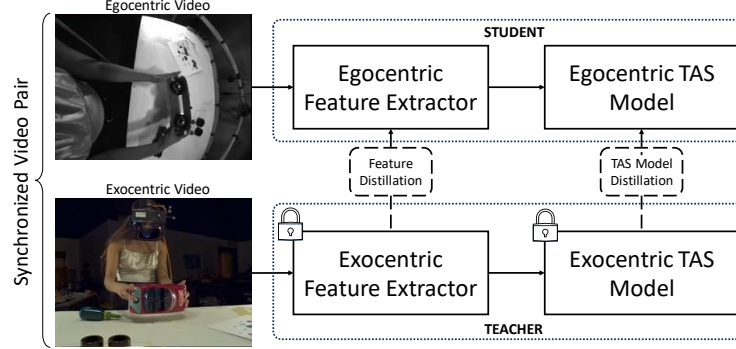
Fig. 2: The proposed method to adapt an exocentric Temporal Action Segmentation model trained on $\mathcal{D}_{train}^{exo}$ to an egocentric setting using a set of unlabelled synchronized video pairs $\mathcal{D}_{adapt}^{pair}$. We investigate distillation at two different levels: the feature extractor and the model. The bottom branch (exocentric) is used to supervise the top branch (egocentric).

Feature Distillation In the feature distillation process, we sample pairs of video segments (S_j^{exo}, S_j^{ego}) from video pairs $(V_j^{exo}, V_j^{ego}) \in \mathcal{D}_{adapt}^{pair}$ and train ϕ^S to minimize the following Mean Squared Error (MSE) loss function:

$$\mathcal{L}_f = \frac{1}{M} \sum_{j=1}^{M} ||\phi^S(S_j^{ego}) - \phi^T(S_j^{exo})||^2 \tag{1}$$

where M is the batch size. In practice, we encourage the student model to extract from the egocentric videos representations similar to the exocentric ones extracted by the teacher model. Additional information about the Feature Distillation process is available in supplementary material.

TAS Model Distillation This process aims to distill the student TAS model γ^S to make its internal representations similar to those extracted by γ^T on the paired exocentric video. Formally, let $\gamma_l(V_j)$ be the features extracted by the prediction model γ at layer l of the architecture when observing egocentric video. We distill the student module by minimizing the following MSE loss function:

$$\mathcal{L}_m = \frac{1}{M} \sum_{j=1}^{M} \sum_{l=1}^{L} ||\gamma_l^S(\phi^S(V_j^{ego})) - \gamma_l^T(\phi^T(V_j^{exo}))||^2 \tag{2}$$

where $\phi(V)$ represents a sequence of frame-wise features densely extracted from video V using the feature extractor ϕ, M is the batch size, L is the number of layers of the model γ and $(V_j^{ego}, V_j^{exo}) \in \mathcal{D}_{adapt}^{pair}$. Additional information about the TAS Model Distillation process is available in supplementary material.
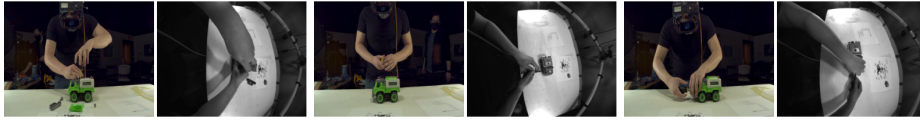
Fig. 3: Three pairs of synchronized exo-ego (v3, e4) frames from Assembly101.

4   Experimental Settings

Assembly101 Dataset We base our main experiments on the Assembly101 dataset [39]. The dataset contains 4321 videos of subjects performing assembly and disassembly procedures of toy models. Videos are collected from multiple synchronized views: 4 first-person monochrome cameras mounted on a custom headset with a resolution of 640×480 pixels, and 8 third-person RGB cameras (5 mounted overhead, 3 on the side) with a resolution of 1920×1080 pixels. The videos contain frame-wise annotations of fine-grained and coarse-grained action annotations. As in [39], we consider coarse-grained action annotations for the temporal action segmentation experiments. For our experiments, we selected two exocentric views (*v3* and *v4*) and one egocentric view (*e4*). This allows us to consider two different adaptation settings $v3 \rightarrow e4$ and $v4 \rightarrow e4$, which allow us to assess the generalization of our method.[5] Each view is associated to 362 videos, which we split as described in the following. We randomly selected 143 exocentric videos to form the exocentric training set $\mathcal{D}_{train}^{exo}$. The corresponding synchronized egocentric videos are included in a $\mathcal{D}_{train}^{ego}$ set. A set of 63 exocentric-egocentric synchronized video pairs is selected to form the adaptation set $\mathcal{D}_{adapt}^{pair}$, whereas 156 egocentric videos are selected to form the egocentric test set $\mathcal{D}_{test}^{ego}$. The corresponding exocentric videos are included in a set $\mathcal{D}_{test}^{exo}$. Examples of exocentric-egocentric video pairs from the dataset are reported in Figure 3. Following the authors [39], we use TSM [29] as the feature extractor ϕ and C2F-TCN [42] as the TAS model. We also experiment with DINOv2 [34] features densely extracted over all video frames. Since fine-tuning DINOv2 is highly demanding in terms of computation, for feature distillation we train an adapter module θ with a residual connection which learns to adapt DINOv2 features extracted by the student model with an MSE loss and use $\theta(\phi(\cdot))$ as a feature extractor at test time. We will release all the details of the proposed split and extracted features to support future research.[6]

EgoExo4D Dataset We also consider the recently proposed EgoExo4D [16] dataset, which contains multi-view videos of subjects performing different activities in diverse scenarios, totaling 1,422 hours of footage. Each video is acquired from an egocentric device and different synchronized GoPro cameras. For each video, we consider the egocentric view and the exocentric one marked as *best view* by humans annotators. We considered the subset of the dataset containing procedural activities, where each video includes hierarchical keystep annota-

---

[5] More information on view selection in supplementary material.

[6] Additional implementation details are in supplementary material.

tions breaking down human activities in temporal video segments. Similarly to Assembly101, we use parent annotations, which are akin to coarse labels. It is worth noting that this dataset provides a much more challenging benchmark as compared to Assembly101, due to the non-standardized placement of cameras, whose placement can vary from a video-to-video basis. To reduce complexity, we selected 5 of the most representative scenarios in the subset of cooking activities: *Making Cucumber and Tomato Salad*, *Cooking Scrambled Eggs*, *Cooking an Omelet*, *Making Sesame-Ginger Asian Salad*, *Cooking Tomato and Eggs*. This leads to 233 videos, which we randomly split in 90 exocentric videos for exocentric training ($\mathcal{D}_{train}^{exo}$), the associated 90 egocentric videos for egocentric training ($\mathcal{D}_{train}^{ego}$), 89 exo-ego video pairs for adaptation ($\mathcal{D}_{adapt}^{pair}$), and 44 exocentric and paired egocentric videos for test ($\mathcal{D}_{test}^{exo}$ and $\mathcal{D}_{test}^{ego}$). We obtain a set of 32 coarse action classes plus a *Background* class, arising from unlabeled parts of the video due to the fact that annotations were not originally designed for temporal action segmentation. To conform to the standard temporal action segmentation scenario in which the background class is generally not present [39], we do not consider this class during training and evaluation. We use C2F-TCN for TAS as in [39] and adopt Omnivore [15] features provided by the authors.[7]

Evaluation Measures All models are evaluated using standard measures, as reported in previous TAS works [10,24,39]. In particular, we report two segment-based evaluation measures, namely edit score and F1 score, and a frame-based evaluation measure, namely Mean over Frames (MoF). All results in this paper are reported in percentage ($0 - 100$ range). See [10] for more information on the definition of the evaluation measures.

5   Results and Discussion

5.1   Performance of the Proposed Approach

Table 1 compares different approaches to exo-to-ego adaptation on the $v3 \rightarrow e4$ view adaptation setting. Exo-oracle baselines (lines 1-2) are trained and tested on exocentric data. These are our teacher models, and are reported here for reference. Ego-oracle baselines (lines 3-4) are trained and tested on egocentric data. They represent the typical results of a supervised model making use of labeled egocentric data. Lines 5-14 report the results of different variations of the proposed method based on knowledge distillation. We compare different choices of backbones and distillation modules.

Oracle Performance We note that exo-oracle methods (lines 1-2) systematically outperform ego-oracle approaches (lines 3-4) according to all evaluation measures when the same backbone is considered (e.g., 31.45 vs 29.25 edit score and 29.92 vs 25.40 $F1_{25}$ comparing line 1 with 3). Note that these approaches have the same architecture but take as input different videos of the same scene (i.e., the egocentric and exocentric views). Hence, the observed difference in

---

[7] https://ego-exo4d-data.org/

Table 1: Comparison of different approaches on the v3 → e4 view adaptation setting. Both feature and TAS model distillation allow to reduce the gap between the baselines (highlighted). Best results within groups are reported in bold. Best results among the adaptation methods are underlined.

	Adaptation Task	Backbone	Feature Dist.	TAS Dist.	Edit	$F1_{10}$	$F1_{25}$	$F1_{50}$	MoF
1	EXO → EXO	TSM			**31.45**	**34.24**	**29.92**	**20.88**	**37.39**
2	(exo-oracle)	DINOv2			28.55	30.13	25.84	17.79	36.03
3	EGO → EGO	TSM			**29.25**	**31.06**	**25.40**	**17.47**	**36.60**
4	(ego-oracle)	DINOv2			26.42	26.50	22.16	14.84	33.47
5					10.53	6.47	2.44	0.56	5.42
6		TSM	✓		12.38	8.73	5.64	1.99	8.21
7				✓	20.17	21.11	18.23	13.12	17.77
8			✓	✓	**25.39**	**26.79**	**22.37**	**15.52**	**30.30**
9	EXO → EGO	Improvement w.r.t. baseline (line 5)			+14.86	+20.32	+19.93	+14.96	+24.88
10					12.60	10.18	7.21	2.40	14.15
11		DINOv2	✓		14.38	10.80	6.53	2.33	10.94
12				✓	<u>**28.59**</u>	<u>**29.58**</u>	<u>**24.84**</u>	<u>**16.38**</u>	31.36
13			✓	✓	28.13	28.75	24.16	15.79	<u>**32.67**</u>
14		Improvement w.r.t. baseline (line 10)			+15.99	+19.40	+17.63	+13.98	+18.52

Table 2: Results for the v4 → e4 view adaptation setting. Best results within groups are reported in bold. Best results among the adaptation methods are underlined.

	Adaptation Task	Backbone	Feature Dist.	TAS Dist.	Edit	$F1_{10}$	$F1_{25}$	$F1_{50}$	MoF
1	EXO → EXO	DINOv2			24.03	19.78	15.17	8.17	22.54
2	EGO → EGO	DINOv2			26.42	26.50	22.16	14.84	33.47
3					12.82	10.88	6.93	2.62	10.65
4		DINOv2	✓		20.24	21.16	17.90	9.90	25.18
5	EXO → EGO			✓	24.23	22.93	18.21	11.91	25.87
6			✓	✓	<u>**24.46**</u>	<u>**25.80**</u>	<u>**21.32**</u>	<u>**13.87**</u>	<u>**29.39**</u>
7		Improvement w.r.t. baseline (line 3)			+11.64	+14.92	+14.39	+11.25	+18.74

performance is likely due to the additional challenges involved with the processing of egocentric video in the Assembly101 dataset, including occlusions, motion blur, and the use of grayscale images (see Figure 3). In this context, the domain-specific TSM model achieves better performance than the domain-agnostic DINOv2 model (e.g., 31.45 vs 28.55 edit score and 20.88 vs 17.79 $F1_{50}$ comparing line 1 with 2 and 29.25 vs 26.42 edit score comparing line 3 with 4). However, remarkably, the gap between DINOv2 and TSM features is small despite DINOv2 has not been trained on any domain-specific videos (e.g., −2.9 edit score comparing line 2 with 1 and −2.83 when comparing line 4 with 3). This highlights the potential of DINOv2 features as a suitable off-the-shelf feature extractor in the considered setup.

Contributions of the Adaptation Modules The proposed adaptation scheme based on knowledge distillation achieves remarkable results both when using domain-specific TSM features (lines 5-9) and when using domain-agnostic DINOv2 features (lines 10-14). For TSM features, we observe consistent improvements when the feature distillation module is activated (comparing lines 6-5, we obtain 12.38 vs 10.53 edit score and 8.21 vs 5.42 MoF). Larger improvements are obtained with TAS model distillation (comparing 7 vs 5-6, we obtain an edit

score of 20.17 vs 10.53 and 12.38). Best results are obtained when both feature distillation and TAS model distillation are activated, with significant +24.88, +20.32 and +14.86 improvements in MoF, $F1_{10}$ and edit score (see line 9).

Similar trends are observed with DINOv2 features, with feature distillation and TAS model distillation bringing improvements with respect to the baseline (compare lines 12-13 vs line 10). Differently from results based on TSM, best overall results are already achieved activating the TAS model distillation, with consistent improvements of +19.40, +17.21 and +15.99 in $F1_{10}$, MoF and edit score respectively (see line 14). Combining both feature and TAS model distillation leads to comparable performance on all evaluation measures, with the exception of MoF which its improved from 31.36 to 32.67 (compare lines 12-13). This marginal improvement suggests the effectiveness of DINOv2 features, which being generic, do not require an explicit adaptation step. Comparing line 10 with line 5, we can observe that the domain-agnostic DINOv2 feature extractor exhibits a smaller exo-ego domain gap (e.g., 12.60 vs 10.53 edit score and 14.15 vs 5.42 MoF). This smaller gap also affects best overall results (numbers in bold), with the best model based on DINOv2 (line 12) obtaining higher performance measures when compared with the best model based on TSM (line 8), e.g., 28.59 vs 25.39 edit score, and 24.84 vs 22.37 $F1_{25}$.

Adaptation Approaches versus Oracle Baselines Interestingly, our best results based on DINOv2 closely match the results obtained by the best ego-oracle baseline (line 12 vs line 3 - e.g., 28.59 vs 29.25 edit score, and 29.58 vs 30.13 $F1_{10}$) and outperform the results obtained by the ego-oracle baseline based on the same DINOv2 feature extractor (line 12 vs line 4 - e.g., 28.59 vs 26.42 edit score, and 24.84 vs 22.16 $F1_{25}$). This highlights that, in the presence of labeled exocentric videos, unlabeled synchronized exo-ego video pairs, and an effective adaptation technique, labeling egocentric videos could be entirely avoided.

Multiple View Transfer Settings Table 2 reports the results on the second view adaptation setting $v4 \rightarrow e4$. Note that, being the target view $e4$ the same as Table 1, results in the two tables are comparable. In this table, we report results only with DINOv2, which achieved best performance in previous experiments. Exo oracle performance highlights that view $v4$ is less informative than $v3$ (compare line 1 of Table 2 with line 2 of Table 1 - e.g. Edit distance 24.03 vs 28.55). Despite this, feature distillation allows to improve over the baseline (compare line 4 with line 3), increasing for instance edit score from 12.82 to 20.24 and MoF from 10.65 to 24.18. The TAS distillation module further improves edit score to 24.23 (line 5), a result closer to the ego oracle (line 2, edit score of 26.42). Enabling both feature and TAS distillation brings additional benefits in particular in the F1 and MoF metrics (compare line 6 with lines 4-5 and line 3), with overall improvements over the baselines of +18.74 in MoF, +14.39 in $F1_{25}$ and +11.64 in edit score. The findings indicate that when the EXO view provides less information, it's preferable to utilize feature distillation combined with TAS model distillation for DINOv2.

Table 3: Comparison of the proposed synchronized model adaptation (last line) with different adaptation approaches. Methods denoted with * are trained on pairs of videos of the same procedure. Best results are in bold, second best results are underlined.

Adaptation	Edit	$F1_{10}$	$F1_{25}$	$F1_{50}$	MoF
No Adaptation	12.60	10.18	7.21	2.40	14.15
MMD [17]	16.98	14.21	10.58	<u>5.86</u>	10.11
DCORAL [46]	17.13	<u>16.25</u>	<u>12.04</u>	5.51	<u>17.72</u>
GRL [14]	16.57	15.43	8.23	2.78	16.84
TCC [11]	16.03	15.11	10.78	5.04	16.42
SDTW [18]	13.82	11.48	8.15	4.77	13.21
TCC [11]*	<u>17.84</u>	15.91	11.59	5.48	17.63
SDTW [18]*	15.27	12.22	9.19	4.94	14.82
Rnd. MSE	13.27	10.03	6.98	2.36	10.61
Sync. (ours)	**28.59**	**29.58**	**24.84**	**16.38**	**31.36**
Imp. vs 2nd best	+10.75	+13.33	+12.8	+10.52	+13.64

5.2 Comparisons with the State of the Art

In Table 3, we compare the proposed method with different state-of-the-art approaches for unsupervised domain adaptation and video alignment. Note that, since no previous investigation has considered the proposed task of exo-to-ego transfer with unlabeled synchronized video pairs, we consider approaches working in the absence of synchronization. These comparisons also serve as a means to assess the contribution of synchronized video pairs to the adaptation process. All these experiments are related to the $v3 \rightarrow e4$ adaptation scenario and use DINOv2 as feature extractor (our best performing configuration), performing TAS model distillation with different loss functions. As domain adaptation approaches, we consider Maximum Mean Discrepancy (MMD) [17], Deep CORAL (DCORAL) [46], and Gradient Reversal Layer (GRL) [14]. As video alignment losses, we consider Temporal Cycle Consistency (TCC) [11] and Smooth Dynamic Time Warping (SDTW) [18]. We compare these approaches with a "No Adaptation" approach, which consists in testing the exocentric model directly on egocentric data, and with a "Rnd. MSE" baseline, which computes a random assignment between video frames and applies the MSE loss to random pairs. When training with the competitor losses, we feed two random exo and ego videos to assess the ability of the model to align features in an unsupervised way. In the case of alignment losses, we also assess the effect of feeding two videos of the same procedure (i.e., an assembly or disassembly of the same object) to ensure that an alignment between the videos indeed exists.

Performing distillation through a random assignment (Rnd. MSE) does not bring systematic improvements on No Adaptation (compare first line with third-last). Domain adaptation approaches (lines 2-4) bring improvements, with DCORAL obtaining an edit score of 17.13 vs 12.60 (+4.53) of No Adaptation. Improvements are less pronounced in the case of unsupervised alignment losses on random video pairs (lines 5-6), with TCC obtaining an edit score of 16.03, a +3.43 over No Adaptation. Considering video pairs of the same procedure sig-

Table 4: Performance with different amounts of unlabeled synchronized Exo-Ego video pairs. The proposed method doubles results over the baseline with 50% of the data, about 5 hours of video.

% Pairs	# Hours	Edit	$F1_{10}$	$F1_{25}$	$F1_{50}$	MoF
0%	0	12.60	10.18	7.21	2.40	14.15
10%	1.04	13.58	9.21	7.78	4.15	10.75
15%	1.43	16.31	15.52	12.4	7.15	17.27
20%	2.09	19.85	18.34	15.31	10.95	21.32
25%	2.78	22.73	23.52	19.29	12.04	26.11
50%	5.23	25.06	25.62	21.07	13.06	29.00
75%	8.06	25.14	26.41	21.87	15.39	29.55
100%	10.21	28.59	29.58	24.84	16.38	31.36

Table 5: DINOv2 model performance with different drop rates. The method shows good performace for small drops, 0.5%, and double baseline performance with a drop of 2.0%

Drop Rate	Edit	$F1_{10}$	$F1_{25}$	$F1_{50}$	MoF
No Drop	28.59	29.58	24.84	16.38	31.36
0.5%	27.68	27.87	23.84	16.29	31.22
1.5%	26.91	27.56	23.86	16.23	30.83
2.0%	26.17	26.38	22.23	14.74	29.69
2.5%	24.63	24.05	20.25	14.13	28.16
3.0%	22.19	21.15	17.98	10.92	25.63
No adaptation	12.60	10.18	7.21	2.40	14.15

nificantly improves results of unsupervised alignment losses (lines 6-7), bringing TCC to the second best result of 17.84 according to the edit score. The proposed distillation approach (second-last line) improves over the second best results by +10.75 in terms of edit score and +13.64 in terms of MoF (see last line). These results highlight the effectiveness of considering unlabeled synchronized video pairs as a source of supervision.

5.3   Amounts of Video Pairs and Accuracy of Synchronization

Table 4 reports the performance of our adaptation method based on DINOv2 when different amounts of unlabeled synchronized Exo-Ego video pairs are used for knowledge distillation. Our model doubles results over the baseline with about 5 hours of video. This highlights the convenience of the proposed approach even in those scenarios in which it is possible to collect only a limited number of synchronized Exo-Ego video pairs.

We further assess the robustness of our method in the presence of imperfect synchronization, which can arise from faulty sensors or unsophisticated synchronization approaches. We trained our DINOv2-based model on $v3 \rightarrow e4$ randomly dropping a given percentage of frames independently from the paired Ego and Exo videos, thus affecting their synchronization. Table 5 reports results for different drop rates, demonstrating that our method is resilient to small drop rates, despite not being explicitly designed for this. For reference, a drop rate of 3% implies that ~ 1 frame/sec is dropped at $30fps$. Even with this significant amount of noise, the proposed adaptation offers advantages over the "no adaptation" baseline. See the supplementary material for additional experiments on training data and feature extractors.

5.4   Analysis on the EgoExo4D Dataset

Table 6 reports results on EgoExo4D. Oracle performance shows that, in this setting, the EGO signal (line 2 Edit 28.74) is stronger than the EXO signal

Table 6: Results on EgoExo4D [16]. 🖊denotes TAS model distillation. Best results within the same group are highlighted.

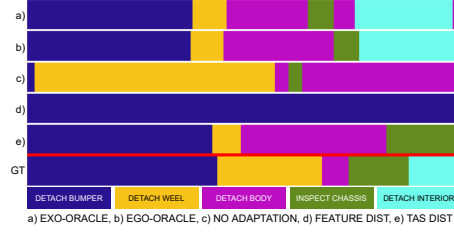|	Adaptation Task	🖊	Edit	$F1_{10}$	$F1_{25}$	$F1_{50}$	MoF
1	EXO → EXO		21.64	16.40	11.69	05.89	24.17
2	EGO → EGO		28.74	24.55	19.88	10.21	40.45
3	EXO → EGO		20.56	16.15	11.20	05.24	24.81
4	EXO → EGO	✓	**23.88**	**18.12**	**12.20**	**05.50**	**30.72**
5	impr. w.r.t. line 3		+3.32	+1.97	+1.00	+0.26	+5.91
6	EGO → EXO		16.05	14.64	10.60	04.82	**26.82**
7	EGO → EXO	✓	**20.49**	**15.62**	**10.98**	**04.90**	25.67
8	impr. w.r.t. line 6		+4.44	+0.98	+0.38	+0.08	−1.15

Fig. 4: Qualitative example of temporal segmentations related to different approaches on a test video.

(line 1 Edit 21.64). We believe that this depends on the high variability in terms of environments and view points of the exocentric cameras which are not stationary across videos. As a result, the gap between a no-adaptation baseline (line 3) and the oracle (line 1) is limited (20.56 vs 21.64). Despite this small gap, the distillation process (line 4) improves results over the baseline (line 3) with improvements of +3.32 in Edit, +5.91 in MoF, and other positive increments in the other measures. Remarkably, the distillation process even surpasses the oracle (23.88 edit in line 4 vs 21.64 in line 1), which is due to the higher quality of information captured by the egocentric view. Given the advantage of the egocentric view in this context, we also experiment with an $EXO \rightarrow EGO$ adaptation, in which an EGO teacher is distilled into a model aiming to improve performance in the exocentric domain. Our methodology generalizes also to this case, with the adapted TAS model (line 7) achieving an edit score of 20.49, an improvement of +4.44 over the no adaptation baseline (line 6), reaching a result comparable to the EXO oracle (line 1 - 21.64 edit).

5.5   Qualitative Results

Figure 4 reports qualitative segmentations of the compared approaches on a test video. Compared methods are based on the DINOv2 feature extractor. More examples and discussion in supplementary material.

6   Conclusion

We have shown that temporal action segmentation models can be transferred from the exocentric to the egocentric domain, provided that labeled exocentric data and unlabeled synchronized exocentric-egocentric video pairs are given. Without bells and whistles, our approach performs on par with fully supervised approaches trained on labeled egocentric data, scoring a +15.99 improvement in the edit score on Assembly101 and +3.32 on the challenging EgoExo4D dataset compared to a baseline trained on exocentric data only. Code is available here: https://github.com/fpv-iplab/synchronization-is-all-you-need.

Acknowledgements

References

1. Camporese, G., Coscia, P., Furnari, A., Farinella, G.M., Ballan, L.: Knowledge distillation for action anticipation via label smoothing. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 3312–3319. IEEE (2021)
2. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
3. Chen, M.H., Li, B., Bao, Y., AlRegib, G.: Action segmentation with mixed temporal domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 605–614 (2020)
4. Chen, X., Qin, S., Bai, Z., Guo, Y., Yang, Z., Jiang, R., Wu, C., Su, S.: Long video-based action segmentation for earthmoving excavators using improved temporal convolutional network models. In: IOP Conference Series: Earth and Environmental Science. vol. 1101, p. 092003. IOP Publishing (2022)
5. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1717–1726 (2020)
6. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 678–695. Springer (2020)
7. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented rgb stream for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7882–7891 (2019)
8. Csurka, G.: Domain adaptation for visual applications: A comprehensive survey. CoRR **abs/1702.05374** (2017), http://arxiv.org/abs/1702.05374
9. Damen, D., Doughty, H., Farinella, G.M., , Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision. CoRR **abs/2006.13256** (2020)
10. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern technique. arXiv preprint arXiv:2210.10352 (2022)
11. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1801–1810 (2019)
12. Fernando, B., Herath, S.: Anticipating human actions by correlating past with the future with jaccard similarity measures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13224–13233 (2021)
13. Furnari, A., Farinella, G.M.: Streaming egocentric action anticipation: An evaluation scheme and approach. Computer Vision and Image Understanding **234**, 103763 (2023)
14. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)

15. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A Single Model for Many Visual Modalities. In: CVPR (2022)
16. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: CVPR (2024)
17. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)
18. Hadji, I., Derpanis, K.G., Jepson, A.D.: Representation learning via global temporal alignment and cycle-consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11068–11077 (2021)
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
20. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer (2017)
21. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning cross-modal contrastive features for video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13618–13627 (2021)
22. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning cross-modal contrastive features for video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13618–13627 (October 2021)
23. Kuehne, H., Richard, A., Gall, J.: A hybrid rnn-hmm approach for weakly supervised temporal action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(4), 765–779 (2020). `https://doi.org/10.1109/TPAMI.2018.2884469`
24. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
25. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10922–10931 (2023)
26. Li, S.J., AbuFarha, Y., Liu, Y., Cheng, M.M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE transactions on pattern analysis and machine intelligence (2020)
27. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: 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 (CVPR). pp. 6943–6953 (June 2021)
28. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European conference on computer vision (ECCV). pp. 619–635 (2018)
29. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
30. Liu, D., Li, Q., Dinh, A., Jiang, T., Shah, M., Xu, C.: Diffusion action segmentation. arXiv preprint arXiv:2303.17959 (2023)
31. Liu, M., Chen, X., Zhang, Y., Li, Y., Rehg, J.M.: Attention distillation for learning video representations. arXiv preprint arXiv:1904.03249 (2019)

32. Liu, X., Zhou, S., Lei, T., Jiang, P., Chen, Z., Lu, H.: First-person video domain adaptation with multi-scene cross-site datasets and attention-based methods. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2023). `https://doi.org/10.1109/TCSVT.2023.3281671`

33. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 122–132 (2020)

34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

35. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2339–2348 (2020)

36. Reza, S., Sundareshan, B., Moghaddam, M., Camps, O.: Enhancing transformer backbone for egocentric video action segmentation. arXiv preprint arXiv:2305.11365 (2023)

37. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)

38. Sayed, S., Ghoddoosian, R., Trivedi, B., Athitsos, V.: A new dataset and approach for timestamp supervised action segmentation using human object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3132–3141 (2023)

39. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)

40. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 154–171. Springer (2020)

41. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7396–7404 (2018)

42. Singhania, D., Rahaman, R., Yao, A.: Coarse to fine multi-resolution temporal convolutional network. arXiv preprint arXiv:2105.10859 (2021)

43. Singhania, D., Rahaman, R., Yao, A.: Iterative contrast-classify for semi-supervised temporal action segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2262–2270 (2022)

44. Spriggs, E.H., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 17–24 (2009). `https://doi.org/10.1109/CVPRW.2009.5204354`

45. Stroud, J., Ross, D., Sun, C., Deng, J., Sukthankar, R.: D3d: Distilled 3d networks for video action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 625–634 (2020)

46. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016)

47. Tran, V., Wang, Y., Zhang, Z., Hoai, M.: Knowledge distillation for human action anticipation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2518–2522. IEEE (2021)
48. Wang, X., Hu, J.F., Lai, J.H., Zhang, J., Zheng, W.S.: Progressive teacher-student learning for early action prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3556–3565 (2019)
49. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(5), 1–46 (2020)
50. Xu, Z., Rawat, Y., Wong, Y., Kankanhalli, M.S., Shah, M.: Don't pour cereal into coffee: Differentiable temporal logic for temporal action segmentation. Advances in Neural Information Processing Systems **35**, 14890–14903 (2022)
51. Xue, Z., Grauman, K.: Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. arXiv preprint arXiv:2306.05526 (2023)
52. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4133–4141 (2017)
53. Yu, H., Cai, M., Liu, Y., Lu, F.: 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. pp. 1358–1366 (2019)