SADA: Semantic adversarial unsupervised domain adaptation for Temporal Action Localization

David Pujol-Perich, Albert Clapés, Sergio Escalera
Universitat de Barcelona and Computer Vision Center, Barcelona, Spain
{david.pujolperich, aclapes, sescalera}@ub.edu

Abstract

Temporal Action Localization (TAL) is a complex task that poses relevant challenges, particularly when attempting to generalize on new – unseen – domains in real-world applications. These scenarios, despite realistic, are often neglected in the literature, exposing these solutions to important performance degradation. In this work, we tackle this issue by introducing, for the first time, an approach for Unsupervised Domain Adaptation (UDA) in sparse TAL, which we refer to as Semantic Adversarial unsupervised Domain Adaptation (SADA). Our contributions are threefold: (1) we pioneer the development of a domain adaptation model that operates on realistic sparse action detection benchmarks; (2) we tackle the limitations of global-distribution alignment techniques by introducing a novel adversarial loss that is sensitive to local class distributions, ensuring finer-grained adaptation; and (3) we present a novel set of benchmarks based on EpicKitchens100 and CharadesEgo, that evaluate multiple domain shifts in a comprehensive manner. Our experiments indicate that SADA improves the adaptation across domains when compared to fully supervised state-of-the-art and alternative UDA methods, attaining a performance boost of up to 6.14% *mAP. The code is publicly available at* https://github.com/davidpujol/SADA.

1. Introduction

Recent advances in the field of video understanding have played a critical role in the surge of novel video-based applications – e.g., video indexing, summarization or recommendation. A critical task of this field is *Temporal Action Localization* (TAL), which involves identifying actions in a video consisting of both their time intervals and action categories. This is particularly difficult given the inherent variabilities of videos. These can be presented, among others, in the form of *appearance variability* – e.g., different kitchens and/or lighting conditions –, *acquisition variability* – e.g., different recording devices – or *viewpoint variability* – e.g., first- or third-person. All these prompt a certain degree of confusion between similar actions to be discriminated.

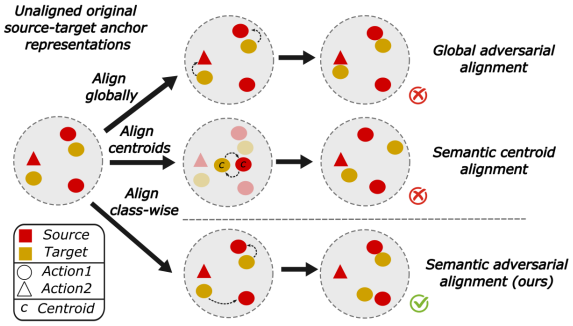


Figure 1. Illustration of the differences between the two most similar domain-adaptation methods [16, 53], and our proposal, *SADA*. For this, we present a simple scenario with various anchor embeddings of different actions (identified by shapes) and domains (identified by colors). In this scenario, [16] (upper row) aligns embeddings in a class-agnostic manner, making it liable to aligning domain embeddings of unmatching action labels. [53] (middle row) computes class-wise mean centroids, and aligns them across domains, but as shown, minimizing their distance does not yield a proper adaptation. *SADA* (last row) improves [16] by aligning class-wise distributions, yielding the correct alignment by not aligning unmatching anchors.

Traditionally, fully supervised methods attempt to address this issue by leveraging enough training data to cover all the possible sources of variability. Unfortunately, this becomes virtually impossible when dealing with realistic scenarios. This compels these methods to operate under the influence of unseen data variations – i.e., *domain gaps* –, which exposes them to a considerable decline in performance. Overcoming this typically involves relabeling data from the new domain, so as to retrain and adapt the model. Unfortunately, this approach is impractical due to the considerable time and resource consumption involved, a challenge exacerbated when dealing with high-dimensional inputs like videos.

Unsupervised Domain Adaptation (UDA) has recently become a hot topic given its potential to leverage unlabelled data to mitigate this domain-induced degradation

[15,20,35]. Despite its considerable success in image-based tasks [10] the application of UDA to video understanding remains underexplored. In fact, to the best of our knowledge, no prior work addresses UDA for TAL setups. The closest proposal is SSTDA [7], which approaches the problem of action segmentation. This focuses on making per-frame action predictions (i.e., *dense TAL*), tackling datasets where no concurrent actions take place [13, 25]. This approach inevitably requires additional smoothing techniques to preserve temporal coherence. This motivates the focus of this paper on the more general *sparse TAL* problem –i.e., segment-level predictions – avoiding the need for additional losses, while intrinsically adapting to multi-label scenarios.

Consequently, in this paper, we propose the first UDA method for sparse multi-label detection on TAL, which we name *Semantic Adversarial unsupervised Domain Adaptation*, or *SADA* for short. Concretely, our proposal specifically builds upon an anchor-based architecture given their recent success on sparse TAL [42, 57]. Thus, our goal is to minimize the discrepancy between anchor representations of a labeled source domain and an unlabelled target domain. These anchors are extracted with a multi-resolution architecture [42] that we couple with a novel adversarial loss that improves the limitations of existing UDA works. Concretely, existing works normally align domain distributions globally [15], applying adversarial methods on the feature embeddings regardless of the action class they represent. As we will show, the coarse alignment of anchors of different action classes with even *background* (no-action) anchors can hurt performance. We propose instead to first use pseudo-labeling [26] to assign an action or background class to each anchor representation. With this, we can factorize the global alignment loss [15] into independent per-class and background distribution alignments. This results in a more sensitive alignment strategy, less prone to *semantic feature misalignment* – i.e., semantically meaningless alignments across domains – and as we will show, better performing.

Assessing the effectiveness of UDA methods for video understanding is a challenging, still unresolved task. Existing proposals on action segmentation [7] follow a subject-based strategy where they aim to adapt a model to new unseen subjects. Here we refer to *subject* as a person appearing in a video. Nevertheless, little data is normally available from a single subject, which inevitably requires grouping several of them for training. This allows the model to generalize over the subject variability under study, making it unsuitable for domain adaptation. To address these limitations, we draw inspiration from the work of [58] on action recognition and investigate the impact of *viewpoint domain shifts* on sparse TAL in CharadesEgo [44]. However, we contend that a more comprehensive evaluation necessitates setups with more controllable shifts. For this, we also propose a suite of 6 new setups based on EpicKitchens100 [12] which study the effect *appearance* and *acquisition* domain shifts. These benchmarks demonstrate that *SADA* mitigates the performance degradation, improving by a large margin the existing fully supervised (namely *source-only*) and UDA-based proposals. In short, our main contributions are:

1. We propose for the first time an UDA method suitable for sparse detection scenarios on TAL.
2. We introduce a novel adversarial loss that factorizes standard global alignment into independent class- and background-wise alignments (see Fig. 1).
3. We present new benchmarks to test sparse detection scenarios when facing 7 different domain shifts, improving the state-of-the-art in all of them.

2. Related work

Temporal Action Localization. At the time of this writing, most of the literature on the task of *Temporal Action Localization* follows a traditional *source-only* approach. In other words, they restrict the models' visibility solely to a training domain, while other domains seen during testing are not available. These works can be categorized as follows: **(1) Anchor-based methods** [2,5,27,28,40,42,46,57] propose a two-stage pipeline, consisting of a proposal generation and classification. The first applies heuristic methods – e.g., uniform sampling [2, 57] or action boundaries' grouping [60, 61] – to generate a dense set of proposals – i.e., temporal segments. In the second stage, they leverage a learnable classifier to predict the corresponding action class and localization offsets of every anchor. Our work falls into this category motivated by the recent success of these methods achieving state-of-the-art results in many TAL benchmarks [12,22,59]. **(2) Anchor-free methods** [28, 29, 43, 56] avoid this two-stage approach making per-frame predictions of their corresponding action labels. These methods, however, often suffer from a tendency towards over-segmentation given the potential discrepancy between neighboring frames. Consequently, they require often complex smoothing techniques to improve the boundary predictions [7]. **(3) Query-based methods** [31, 32, 47] recently emerged as an alternative paradigm that follows the principles presented by [3]. This approach exploits the use of a Transformer encoder-decoder architecture [50] to learn a fixed small set of queries given refined video features, each identifying one potential action segment. Intuitively, this results in a non-heuristic-based proposal generation. This comes with the limitation of an increased rigidity, as the number of proposals needs to be fixed beforehand.

Unsupervised Domain Adaptation. Domain Adaptation techniques emerge as an effective solution to bridge the gap between data collected from a source and a target distribution, respectively. A large suite of approaches has been proposed to perform this alignment between labeled and un-

labeled domains – e.g., discrepancy minimization [24, 54] or entropy minimization [18, 55]. Arguably, nowadays the most popular approach is based on adversarial training [15, 17, 20, 21, 49]. These learn domain-invariant embeddings [11] by training in a min-max fashion a domain classifier to discern if samples come from the source or the target domain. Despite convenient, the simplicity of these methods often degrade the quality of the alignment [26], as they potentially align embeddings of source and target domain that represent different semantic information – e.g., different class labels. Few works have been proposed to do this alignment in a more sensitive way [26]. Works like [21, 53], for instance, reduce the distance of per-class centroids, normally computed as the mean feature embeddings of a given class. Its effectiveness, however, relies on the assumption that the data is distributed somewhat homogeneously around the center, as otherwise, the centroids are not necessarily meaningful. In our work, we couple the advantages of both adversarial domain adaptation and semantic alignment and propose for the first time a pure adversarial semantic loss that yields domain invariant representations in a semantically meaningful way, without making explicit assumptions of the distributions (see Fig. 1).

Unsupervised Domain Adaptation for TAL. Despite the considerable success of UDA methods, their applicability has been mostly restricted to image-based scenarios such as image classification [15, 19, 33, 35] or object detection [8, 37]. Much less attention has been dedicated to video-based applications such as action recognition [6, 23, 38] or spatio-temporal action segmentation [1, 34]. To the best of our knowledge, at the time of this writing, there is no direct comparison with our work focusing on UDA for *sparse* TAL. The closest work is SSTDA [7] that applies UDA for Action Segmentation. SSTDA proposes the use of two global-distribution-based auxiliary tasks to jointly align cross-domain feature spaces. Unlike our proposal, their work falls into the category of anchor-free, making per-frame action predictions. This restricts its applicability to action segmentation scenarios, where current datasets [13, 25] are designed to deal with frame-based single-action classification. In our work, we overcome this limitation by leveraging an anchor-based architecture that enables a natural adaptation to more realistic multi-label scenarios.

3. Method

3.1. Problem definition and notation

In this paper, we address the problem of unsupervised domain adaptation for TAL. For this, we define a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. Domain $\mathcal{S}$ consists of $N_{\mathcal{S}}$ labeled input videos $\{(V_k^{\mathcal{S}}, Y_k^{\mathcal{S}})\}_{k=1}^{N_{\mathcal{S}}}$, where each video $V_k^{\mathcal{S}}$ is a sequence of T frames $(X_{k,1}, \dots, X_{k,T})$ with $X_{k,t} \in \mathbb{R}^{H \times W \times C}$. Here $Y_k = \{(b_{k,i}, e_{k,i}, c_{k,i})\}_{i=1}^{G_k}$ contains the begin, end, and class actions of all the ground-truth (GT) segments G_k of the k-th video, respectively. The target domain $\mathcal{T}$ is similar to $\mathcal{S}$ but lacks the GT information. Concretely, it consists of $N_{\mathcal{T}}$ unlabeled input videos $\{V_k\}_{k=1}^{N_{\mathcal{T}}}$. Our goal is to train a model that can identify the action segments, including both segment coordinates and action labels, in videos from domain $\mathcal{S}$, while minimizing the performance degradation on the unlabeled domain $\mathcal{T}$.

3.2. Framework overview

We propose a model based on a feature pyramid and an anchor-based classification and localization head (see Fig. 2). This architecture is coupled with a novel *semantic adversarial loss* that aligns the anchor embeddings across domains $\mathcal{S}$ and $\mathcal{T}$ in a semantically meaningful way. More in detail, the model takes as input two videos from domain $\mathcal{S}$ and $\mathcal{T}$, respectively. The model first processes the raw input videos using a frozen pre-trained video backbone. The resulting embeddings of both domains are then passed through a shared SGP pyramid [42] that outputs a set of multi-resolution anchor embeddings for each of the predefined resolution levels. The main goal of our model is to make these anchor embeddings domain invariant. For this, we introduce a level-wise *semantic adversarial loss* that learns in an adversarial manner to align the embeddings of both domains belonging to a given action class i at a given resolution level l. Recall that GT information is only available for domain $\mathcal{S}$, therefore we rely on the use of pseudo labeling techniques [26] to infer the *probable* class labels of the data from domain $\mathcal{T}$. Finally, we use the domain invariant anchors of domain $\mathcal{S}$ to train a classification and localization head that learns the underlying tasks in a standard supervised fashion. In short, this permits to learn a classification and localization head that minimizes the decline of performance when applied to the unseen domain $\mathcal{T}$.

3.3. Backbone and SGP pyramid

Our model first takes two input videos $V^{\mathcal{S}}$ and $V^{\mathcal{T}}$ of both domains $\mathcal{S}$ and $\mathcal{T}$. For simplicity, both videos $V^{\mathcal{S}}$ and $V^{\mathcal{T}}$ have length T, which we enforce using padding. The method then processes the two videos applying a frozen pre-trained backbone – e.g., I3D [4] or Slowfast [14]. This permits to extract, in an effective way, temporal cues of the video into a set of refined video features. These embeddings are then fed to an SGP feature pyramid [42] which combines the use of SGP blocks and the progressive downsampling of the temporal length by a ratio of 2. This outputs a set of multi-resolution anchor embeddings $Z^{\mathcal{S}} = \{Z_l^{\mathcal{S}}\}_{l \in L}$ and $Z^{\mathcal{T}} = \{Z_l^{\mathcal{T}}\}_{l \in L}$, for the two domains, respectively. Here L denotes the set of predefined resolution levels and $Z_l \in \mathbb{R}^{T_l \times F}$ the anchor embeddings of level l of a given domain. Concretely, this permits to obtain embeddings for a set of uniformly sampled anchors at each of the $l \in L$

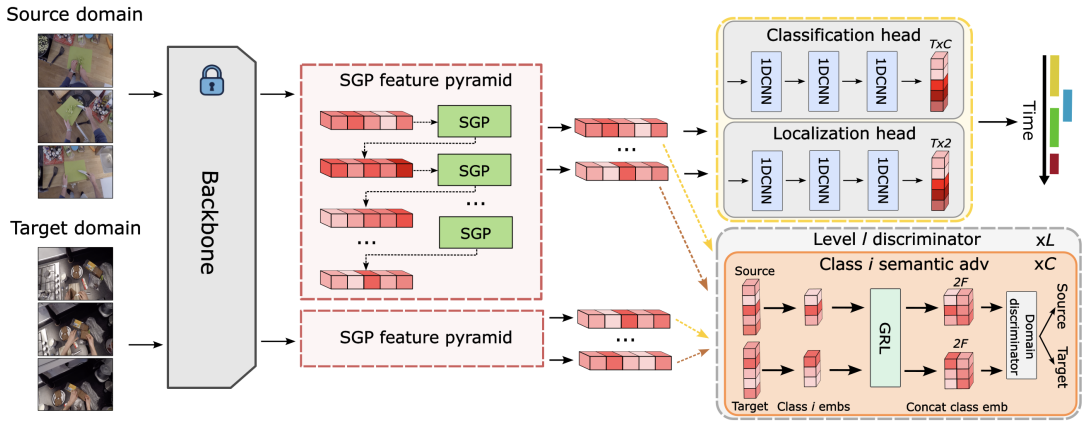


Figure 2. Overview of the main model architecture of SADA. This takes as input videos from a Source and a Target domain, which are both fed to a shared multi-resolution feature extractor pyramid. The output embeddings of both of these domains are then aligned using the semantic alignment loss, *SADA*. This is done with a level and class-wise domain discriminator of the filtered embeddings, based on GT information and pseudo labels, for the source and target domains, respectively. Finally, the resulting domain invariant representations of the source domain are used to train a classification and localization head to learn the underlying task.

resolution levels. The use of a multi-resolution model is favorable for naturally adapting to different action lengths and abstraction levels.

3.4. Classification and localization head

To learn the underlying TAL task, we train in a fully supervised manner a classification and localization module with the labeled source domain $\mathcal{S}$. Due to the anchor-based nature of our model, we first require a matching strategy between the set of candidate anchors to the actual GT segments. For this, we follow a center sampling strategy [42,48]. In other words, for a given level l, we define an anchor as *action anchor* if the time instant t that it represents is near the center of an action. All the rest are marked as *background anchors*. We define $\mathcal{B}_l, \mathcal{E}_l$ and $\mathcal{C}_l$ as the begins, ends and action classes of their matching GT segments. We identify *background anchors* with action label 0. With this, we design a classification head $H_{cls} : \mathbb{R}^{T_l \times F} \rightarrow \mathbb{R}^{T_l \times C}$ that maps each of the anchor embeddings to their class distribution. More specifically, we model this as a sequence of 1D convolutions, and train it using a sigmoid focal loss [30]:

$$\mathcal{L}^l_{SFL} = SFL(H_{cls}(Z^{\mathcal{S}}_l),\ \mathcal{C}_l). \tag{1}$$

Similarly, we model a localization head $H_{loc} : \mathbb{R}^{T_l \times F} \rightarrow \mathbb{R}^{T_l \times 2}$ identically as H_{cls}, which predicts the begin-end offsets. We thus define the localization loss as a standard mean squared error (MSE) loss over the *action anchors* only:

$$\mathcal{L}^l_{loc} = MSE(H_{loc}(Z^{\mathcal{S}}_{l_+}),\ (\mathcal{B}_{l_+} \,||\, \mathcal{E}_{l_+})), \tag{2}$$

where l_+ refers to the filtered *action-anchor* representations from the GT of the l-th level only, and $||$ to the concatenation operation. This yields the final task loss defined as:

$$\mathcal{L}_{task} = \lambda_{cls} \sum_{l \in L} \mathcal{L}^l_{SFL} + \lambda_{loc} \sum_{l \in L} \mathcal{L}^l_{loc}. \tag{3}$$

Here λ_{cls} and λ_{loc} are two tunable hyperparameters.

3.5. Our proposal: Semantic adversarial multi-resolution alignment

One of the main contributions of this paper is the design of a novel adversarial-based loss that we name *SADA* loss, which attempts to overcome the limitations of the extensively used *global adversarial loss* [15]. Traditional adversarial domain adaptation relies on the idea of designing a domain classifier that learns to identify the domain that each of the embeddings belongs to. The rest of the model learns concurrently the opposite objective which results in the learning of domain invariant representations [15]. While this approach has been shown to be effective in other fields – e.g., image classification or object detection – we find that its performance in more challenging video understanding setups like TAL presents important challenges. One of the main issues, as argued by [26], is that this loss often suffers from *feature misalignment* which greatly declines its effectiveness. This refers to the cases where these methods align embeddings of non-matching class labels – i.e., aligning embeddings of domain $\mathcal{S}$ of an action i with embeddings of domain $\mathcal{T}$ of a class j. This issue is further exacerbated in TAL given the noise induced by the alignment of the many *background anchors* with the *action anchors*. We study this phenomenon in more depth in the Supp.

Local adversarial alignment: To fix this *feature misalignment* in realistic scenarios like TAL, we propose an alternative adversarial loss formulation that provides a finer-grained alignment. This loss first attempts to perform a local class-aware alignment. This is, for every given resolution level l, we group the *action anchors* – those matched with a GT action – of $\mathcal{S}$ and $\mathcal{T}$ according to their action label i. This is straightforward for domain $\mathcal{S}$ as we have the GT information. In contrast, for domain $\mathcal{T}$, we use a hard-pseudo labeling strategy [26] that classifies a given embed-

ding as class i if this is the highest-confidence score of the predicted class distribution, and this is above a threshold α. Formally we define the pseudo-label of an anchor z as:

$$\hat{c}_z = \begin{cases} \underset{i}{\text{argmax}}\ P_l[z,i] & \text{if } P_l[z,i] > \alpha \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $P_l = H_{cls}(Z_l^{\mathcal{T}}) \in \mathbb{R}^{T_l \times C}$ are the predicted class probabilities of the anchors. Notice that we mark with class 0 the *background anchors*, which are not assigned to any action class. From this, we obtain the newly grouped embeddings of source and target domain of class i on level l

$$A_i^l = \{Z_l^{\mathcal{S}}[z] \ : \ c_z = i\}_{z \in T_l}, \tag{5}$$

$$B_i^l = \{Z_l^{\mathcal{T}}[z] \ : \ \hat{c}_z = i\}_{z \in T_l}, \tag{6}$$

for $A_i^l \in \mathbb{R}^{T_{l,i} \times F}$, $B_i^l \in \mathbb{R}^{T_{l,i} \times F}$. Also, c_z is the GT action label of anchor z and $\hat{c}_z$ is its computed pseudo-label from Eq. 4. We then adversarially train a single domain classifier $D : \mathbb{R}^{2F} \rightarrow \{0, 1\}$ to identify the domain of each of these embeddings using a binary cross entropy (BCE) loss:

$$\mathcal{L}_{local}^l = \sum_{i=1}^{C} (\mathcal{L}_{\text{BCE}}(D(A_i^l \,||\, E_i), d_{\mathcal{S}})) + (\mathcal{L}_{\text{BCE}}(D(B_i^l \,||\, E_i), d_{\mathcal{T}})), \tag{7}$$

where $d_{\mathcal{S}}$ and $d_{\mathcal{T}}$ are the domain labels. We then introduce a Reverse Gradient Layer (GRL) [15] before the discriminator D to invert the gradients sign, creating a min-max game where the feature extractor learns to *confuse* the discriminator. We condition the discriminator to class i using a learnable class embedding $e_i \in \mathbb{R}^F$ that we replicate for every selected anchor into an embedding E_i (see Supp. for an ablation of this model decision).

Local and global alignment (*SADA*): Eq. 7 aims solely to align the *action anchors* – which are classified as one of the C classes – but *what happens with the background embeddings that fall below the threshold α?* In this case, the loss ignores their influence, yielding only partial alignment.

To overcome this issue, we propose our final *SADA* loss which attempts to combine the best of both *global alignment loss* [15] and Eq. 7. For this, we introduce a new loss term for the *background anchors* as follows:

$$\mathcal{L}_{bkg}^l = \mathcal{L}_{\text{BCE}}(D(A_0^l \,||\, E_0), d_{\mathcal{S}}) + \mathcal{L}_{\text{BCE}}(D(B_0^l \,||\, E_0), d_{\mathcal{T}}), \tag{8}$$

where again A_0^l and B_0^l are the selected *background anchors*, and $E_0 \in \mathbb{R}^F$ is the learnable *background* embedding. Coupling Eq. 7 and Eq. 8 yields the final formulation of our proposed loss, combining *local* (class-wise) alignment with the *background* anchors alignment. Formally:

$$\mathcal{L}_{sada} = \sum_{l \in L} \lambda_l \ (\mathcal{L}_{local}^l + \mathcal{L}_{bkg}^l), \tag{9}$$

where λ_l is a hyper-parameter that modulates the importance of level l on the final alignment loss. See Supp. for an analysis of our model's sensitivity to this parameter choice.

3.6. Training

During training, we formulate the final loss as a min-max game where the main model architecture is optimized over the classification and localization loss while maximizing the adversarial loss. In parallel, the discriminator model D attempts to minimize the discriminator loss only. Formally,

$$\mathcal{L} = \lambda_{task}\mathcal{L}_{task} + \lambda_{sada}\mathcal{L}_{sada}. \tag{10}$$

Note again $\mathcal{L}_{task}$ is optimized with domain $\mathcal{S}$ while $\mathcal{L}_{sada}$ promotes the alignment between both domains $\mathcal{S}$ and $\mathcal{T}$. Moreover, λ_{task} and λ_{sada} are tunable parameters.

4. Datasets and experiments

In this section, we present a novel comprehensive benchmark to evaluate the task of UDA for TAL. Our setup, for the first time, goes beyond action segmentation and evaluates the adaptation to different domain shifts in more realistic scenarios with sparse multi-label annotations. We then showcase the effectiveness of our model over the state-of-the-art methods together with several relevant ablations.

4.1. Benchmarks for UDA on sparse TAL

Evaluating domain adaptation-based methods in the context of video understanding is a challenging issue that requires the definition of a reasonable domain gap and identifying a sufficiently large set of intersecting action classes. Dividing existing datasets into different domains that comply with these conditions often restricts the amount of data to learn and adapt. SSTDA [7] approaches this problem on GTEA [13] and Breakfast [25] by defining a subject-based partitioning where they aim to adapt the model to new unseen subjects. However, as little data is available from a single subject in those datasets, they group several users for training. This allows the model to generalize over the subject variability under study, making it unsuitable to test for domain adaptation. Closely related to our work, [36, 51] propose several UDA scenarios for video classification based on EpicKitchens100 [12]. These define 3 domains based on the data of 3 different kitchens, thus performing cross-kitchen evaluation. This limits even further the amount of data in each domain – i.e., between 15 to 29 videos per domain.

EpicKitchens100: To overcome these limitations, we first propose a new set of 6 different scenarios *(S1, ..., S6)* for sparse TAL based on EK100 [12], (see Fig. 3). EK100 presents an ideal base for our tasks as it has become a gold standard to evaluate complex sparse detection scenarios on long egocentric videos (up to 45 minutes). We identify two domain gaps in this dataset: an *appearance domain shift* based on the different colors of the kitchen counters; and an *acquisition domain shift* that results in the differences of lighting and camera conditions when extending EK55 [12]

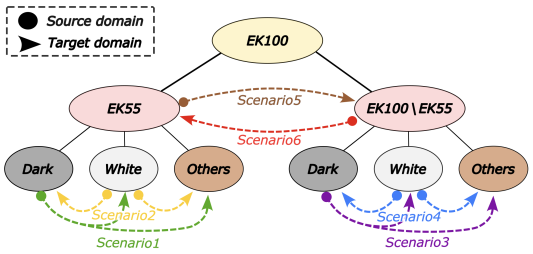


Figure 3. Overview of the 6 proposed experimental setups for EpicKitchens100. Concretely, S1 and S2 evaluate the videos from the original EK55. They define the dark-counter and white-counter kitchens as Source, respectively, and the rest as Target. S3 and S4 are similar except that they consider only the *newest* videos from EK100. S5 and S6 use the *old* videos as Source and the *new* videos as Target, and vice versa.

into its new version EK100 [12]. This permits to define a rich set of benchmarks that provide a more fine-grained and comprehensive evaluation. Importantly, this strategy is also suitable for single-source settings which do not allow an easy generalization over the shift under study. For example, if a model is trained with dark-counter kitchens only, it cannot easily generalize to white-counter kitchens.

More in detail, we introduce 4 different benchmarks that derive from the aforementioned *appearance shift*. For this, we first define two different splits of *EK55* and $\{$*EK100* $\backslash$*EK55*$\}$ videos according to [12], put more plainly *old* versus *new* videos. We then further split each of them according to the color of the kitchen counter and define 4 different setups of single-source domain and multi-target domain – i.e., *dark* $\rightarrow$ *white* and *other* kitchens; and *white* $\rightarrow$ *dark* and *other* kitchens, for *old* and *new* videos, respectively. Similarly, given the two splits between *EK55* and $\{$*EK100* $\backslash$*EK55*$\}$ videos, we propose 2 additional setups of *EK55* $\rightarrow$ $\{$*EK100* $\backslash$*EK55*$\}$ and $\{$*EK100* $\backslash$*EK55*$\}$ $\rightarrow$ *EK55*. These measure the adaptability to different acquisition conditions – i.e., lighting and camera conditions. This results in domains that contain in the order of hundreds of videos. See Supp. for the domain visualizations and statistics.

One important consideration is that EK100 exhibits a very strong long-tail label distribution, where the vast majority of its 97 actions represent only a marginal percentage of the overall data (see Supp.). Adapting DA methods to these long tail distributions falls beyond the scope of this paper and most of the existing literature. For this reason, we follow a similar approach to [36] and limit our evaluation to the 10 majority classes for both domains $\mathcal{S}$ and $\mathcal{T}$, representing 80% of the original data. This still results in a very complex task, proof of which is that the best-performing models to date attain less than a 30% mAP [42, 57].

CharadesEgo: One important aspect of the 6 previous setups is that all the domains share an egocentric viewpoint. For this reason, we evaluate the model degradation under extreme domain shifts caused by major changes in the perspective. For this, we use CharadesEgo [44] which extends the original third-person videos from Charades [45] by matching them with their corresponding egocentric videos. Following [58], we define the Source domain as the third-person videos, and the Target domain as the egocentric ones. Given the extreme shift that these present, we limit the task to predicting the verb of each of the action segments, and similarly to the previous case, we keep the 10 majority verbs only. We refer to the Supp. for more details.

4.2. Experimental results

In this section, we present the main experimental results evaluating *appearance* and *acquisition shifts*. All these experiments follow the standard transductive unsupervised DA protocol [9, 39], and report the mean average precision (mAP) at different intersections over union (IOU) thresholds ($10\% - 50\%$).

EpicKitchens100. In Tab. 1 we first present the results obtained in the 4 different scenarios that we designed to evaluate the performance of our method when facing different *appearance shifts* induced by changes in the background information. Concretely, we first evaluate the performance of the Actionformer [57] and TriDet [42], the two best-performing methods on EK100 dataset [12], as well as our proposed architecture without our SADA loss. We then compare these 3 architectures with their extensions which include the *SADA* loss. Observe that our proposed loss improves the source-only (SO) version of all the architectures in all these 4 scenarios, showing robustness across the chosen underlying architecture. For instance on S3, our proposed loss yields an improvement of up to 2.49% mAP over its respective SO version. Additionally, we observe that our final architecture – i.e., Ours(SADA) – improves the best-performing existing SO method – i.e., Tridet [42] – by up to a 3.4% mAP for the *black-counter kitchens* scenarios – i.e., S1 and S3 – and 1.82% mAP for the *white-counter kitchens* –i.e., S2 and S4. Similarly, the last two scenarios of Tab. 1 present a similar behavior than before. In all but one case on the Tridet [42] architecture, the use of the SADA loss yields a performance gain of up to 2.28% mAP. This is a relative 12.48% improvement. Our final model, moreover, improves the best-performing existing baseline – i.e., Tridet – by 1.73% mAP and 1.25% mAP, respectively.

CharadesEgo. In Tab. 2 we report a similar experimental comparison of our proposed loss when evaluated on CharadesEgo. Concretely, we showcase the performance boost that SADA reports on the three test architectures – i.e., Actionformer [57], Tridet [42] and Ours. Observe that SADA, for instance, improves the performance of Tridet [42] by 1.37% mAP. Similarly, our proposed architecture *Ours (SADA)* improves the SO version by 0.75% mAP and yields the overall highest scores.

Scenario	Model	mAP {10,20,30,40,50}%					Avg	Scenario	Model	mAP {10,20,30,40,50}%					Avg
S1	Actionformer [57]	30.21	28.73	26.39	22.60	17.09	25.00	S2	Actionformer [57]	27.46	26.54	24.61	21.85	17.19	23.53
	Tridet [42]	29.87	28.39	25.97	22.06	16.94	24.65		Tridet [42]	30.03	28.97	26.96	23.48	18.18	25.52
	Ours (src-only)	30.74	29.47	27.23	23.53	18.07	25.80		Ours (src-only)	29.65	28.69	26.86	23.88	19.14	25.64
	Actionformer+SADA	31.71	30.30	27.95	24.08	18.69	26.55		Actionformer+SADA	29.84	28.89	26.83	23.76	18.85	25.63
	Tridet+SADA	29.55	28.27	26.16	22.89	17.63	24.90		Tridet+SADA	30.21	29.33	27.58	24.35	19.44	26.18
	Ours (SADA)	**31.60**	**30.29**	**28.22**	**24.47**	**18.98**	**26.72**		**Ours (SADA)**	**31.54**	**30.68**	**28.77**	**25.52**	**20.22**	**27.34**
S3	Actionformer [57]	28.11	26.94	24.89	21.43	16.51	23.57	S4	Actionformer [57]	33.52	32.31	29.84	26.48	20.11	28.45
	Tridet [42]	29.47	28.32	25.50	21.99	16.34	24.32		Tridet [42]	34.01	32.52	30.07	26.40	19.78	28.55
	Ours (src-only)	30.03	28.70	26.62	23.03	17.79	25.23		Ours (src-only)	34.41	33.38	30.58	26.99	21.00	29.27
	Actionformer+SADA	30.54	29.31	27.36	23.75	18.80	25.95		Actionformer+SADA	34.11	32.87	30.60	27.05	20.93	29.11
	Tridet+SADA	31.34	30.16	27.94	24.48	19.31	26.64		Tridet+SADA	34.50	33.19	30.65	27.25	20.83	29.29
	Ours (SADA)	**32.69**	**31.49**	**29.17**	**25.51**	**19.72**	**27.72**		**Ours (SADA)**	**34.86**	**33.73**	**31.16**	**27.45**	**21.46**	**29.73**
S5	Actionformer [57]	22.87	21.87	20.10	17.23	13.33	19.08	S6	Actionformer [57]	22.16	21.22	19.71	17.44	14.08	18.92
	Tridet [42]	24.77	22.93	21.49	19.09	15.15	20.48		Tridet [42]	22.47	21.57	20.19	17.87	14.41	19.30
	Ours (src-only)	25.58	24.79	23.08	19.56	15.15	21.63		Ours (src-only)	20.96	20.22	19.08	16.97	14.09	18.27
	Actionformer+SADA	24.85	23.99	22.21	19.11	15.28	21.09		Actionformer+SADA	23.05	22.10	20.71	18.31	14.72	19.78
	Tridet+SADA	24.88	24.00	22.20	19.02	14.88	21.00		Tridet+SADA	21.00	20.08	18.64	16.39	13.13	17.85
	Ours (SADA)	**25.93**	**25.06**	**23.47**	**20.45**	**16.12**	**22.21**		**Ours (SADA)**	**23.94**	**22.95**	**21.47**	**19.16**	**15.24**	**20.55**

Table 1. Comparison with SOTA for the 4 appearance-shift scenarios (1-4) and the 2 acquisition-shift scenarios (5-6) on EpicKitchens100.

Model	mAP {10,20,30,40,50}%					Avg
Actionformer [57]	31.22	28.51	23.82	18.91	13.94	23.28
Tridet [42]	30.18	27.06	22.58	17.81	13.10	22.15
Ours (src-only)	30.68	27.74	23.41	18.46	13.58	22.77
Actionformer+SADA	31.46	28.57	**24.17**	19.04	13.87	23.42
Tridet+SADA	31.53	28.41	24.04	18.88	13.98	23.37
Ours (SADA)	**31.68**	**28.64**	24.09	**19.06**	**14.14**	**23.52**

Table 2. Comparison with the state-of-the-art on CharadesEgo.

Model	mAP {10,30,50}%			Avg
DANN [16]	30.48	27.07	18.12	25.22
ADDA [49]	31.84	28.53	19.11	26.49
WDGRL [41]	25.05	22.08	13.91	20.35
FGDA [17]	26.21	22.99	14.12	21.10
DRDA [21]	30.79	26.97	18.60	25.45
MSTN [53]	31.07	27.16	17.21	25.14
SSTDA [7]	31.17	28.01	18.98	26.05
TranSVAE [52]	29.91	26.12	16.16	24.06
Ours (SADA)	**32.69**	**29.17**	**19.72**	**27.19**

Table 3. Comparison with SOTA UDA methods on S3.

4.3. Ablation studies

Comparing to other domain adaptation methods. In Sec. 4.2 we showed that *SADA* consistently improves the performance of the three tested state-of-the-art SO architectures when evaluated on our newly proposed setups. The question remains however of how well does *SADA* perform compared to existing domain adaptation methods. In this regard, following existing video-based domain adaptation works, we first evaluate various canonical UDA domain adaptation methods –i.e., DANN [16], ADDA [49], WDGRL [41], MSTN [53], FGDA [17] and DRDA [21]. These methods are integrated into our proposed underlying architecture resulting in a fair comparison with *SADA*. Additionally, we find that to the best of our knowledge, there is no existing UDA method for TAL in the literature that is directly comparable to us. Nevertheless, to provide a richer comparison, we adapt the closest action segmentation proposal, SSTDA [7] and a state-of-the-art domain-adaptation method for video classification – i.e., TranSVAE [51] – to our proposed setup (we refer to the Supp. for all the details of all these baselines). Concretely, in Tab. 3 we compare *SADA* with these baselines on S3. These results empirically demonstrate the effectiveness of our method by attaining the best results over all tested UDA methods. More in detail, *SADA* yields an improvement of up to 0.7% mAP over the second best-performing method (ADDA). Find the complete ablation for the other scenarios in the Supp.

Analysis of our loss. Next, we ablate over different variants of our proposed *SADA* loss (see Eq. 9). Concretely, in Tab. 4 we study the effect of class-wise distribution alignment (Eq. 7), global distribution alignment (equivalent to DANN [16]) and semantic background alignment (Eq. 8). In this regard, we highlight that the global adaptation seems to consistently improve upon aligning only background embeddings. This is because the latter yields only a partial alignment of the embeddings, not considering *class anchors*. Therefore, a *rougher* yet complete adaptation might seem beneficial. We also observe that aligning local class-wise distributions has a considerable positive effect. Its effect, however, considerably decreases when combined with a global alignment [16], as all the *class anchors* are then subject to the concurrent alignments of domain-level and class-wise distributions. This observation is also consistent with the performance decrease that we observe when combining global and background alignment, which again suggests that the concurrent alignment of background embeddings with two adaptation losses is harmful to performance. Finally, we observe that we consistently obtain the best results when the local alignment loss is coupled with the complementary (non-overlapping) background loss – i.e., SADA loss – indicating that this semantic fine-grained, yet complete, alignment is the most desirable approach. Find in the Supp. the complete ablations for additional scenarios and the class-wise analysis.

Analysis of the impact of *background anchors*. One critical aspect of anchor-based methods for TAL is the presence of numerous *background anchors*. We argue that the confusion that these embeddings induce is one of the main chal-

Local	Global	Bkg	mAP {10,30,50}%			Avg
			30.03	26.62	17.79	24.81
	✓		30.48	27.07	18.12	25.22
		✓	30.34	26.66	17.04	24.68
	✓	✓	29.76	26.63	17.52	24.64
✓			31.36	28.01	18.67	26.01
✓	✓		30.09	26.88	17.43	24.80
✓		✓	**32.69**	**29.17**	**19.72**	**27.19**

Table 4. Ablation of the effect of the components of *SADA* on S3.

Method	Mask *bkg anchors*	mAP {10,30,50}%			Avg	Perf. gap
Ours(src-only)		30.03	26.62	17.79	24.81	-
	✓	35.25	33.71	23.25	30.73	5.92
Ours(SADA)		32.69	29.17	19.72	27.19	-
	✓	**35.70**	**34.09**	**24.15**	**31.31**	**4.12**

Table 5. Ablation of the effect of the *background anchors* in the performance of the model on S3.

lenges for the transfer to a different domain. Concretely, we hypothesize this is because of their high intra-class variance, and their low inter-class variance with other action classes –i.e., they share aspects like appearance. Nevertheless, in this ablation we show that SADA partially mitigates this effect. For this, we design an experiment that artificially masks out all the *background anchors* during inference to elucidate the ideal performance if we could ignore entirely the confusion they generate –i.e., by being wrongly predicted as an action. Concretely, in Tab. 5 we compare our proposed methods with its SO version, and observe that masking out all the *background anchors* yields an overall improvement of 4.12% and 5.92% mAP, respectively. Hence, SADA reduces the impact of masking the *background anchors* by 1.8% mAP, indicating that our fine-grained alignment permits a better knowledge transfer to the target domain, mitigating the negative effect of these anchors. Find in the Supp. the complete ablation.

Qualitative analysis. We complement our quantitative results with a qualitative study. For this, we depict in Fig. 4 a segment visualization of S3 of our proposed method versus the chosen baseline models. In this visualization, we can observe that the Actionformer misses many of the segments in the shown video clip ignoring all the *take* actions and mistaking a *close* for a *take*. Tridet performs better but misses all the *take* actions in the first half of the video while worsening the boundary prediction of the last *close*. *SADA* improves Tridet by predicting the second *take* and considerably improving the boundary of the last *close* action.

Fig. 5 also shows the TSNE plots of the domain-invariant embeddings of *SADA* with respect to learning source-only. Given that *SADA* is an adversarial-based class-wise loss, we depict the plots of the 3 majority action classes (first 3 columns). Observe that SO (top row) yields clearly unaligned distributions with scarce to no overlap in the projected space. Our method, in contrast, presents a considerable distribution mix improving the alignment of class-wise distributions across domains. Given the anchor-based nature of our method, we have numerous *background anchors* –i.e., not assigned to any GT label. As observed in the last column, these are also aligned by our method (see Eq. 8) therefore effectively aligning the entire data distributions but in a semantically sensitive way. See the Supp. for the complete study at different resolution levels.

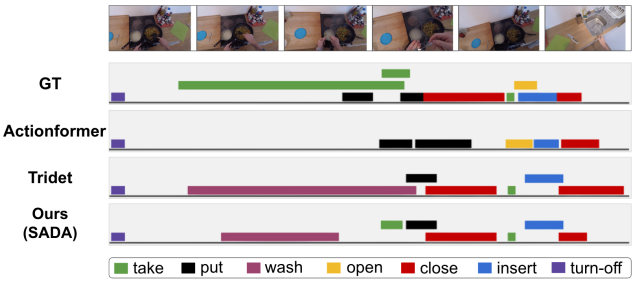
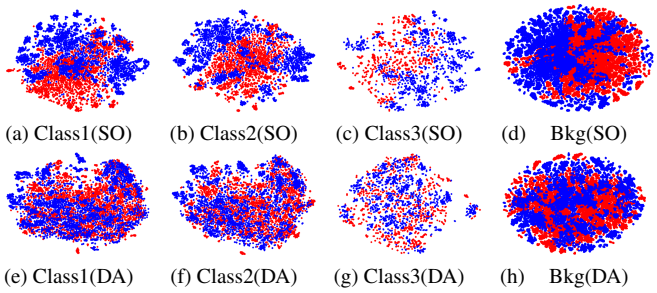


Figure 4. Visualization of the predicted segments of our method and the chosen set of source-only (SO). We include on top the ground-truth (GT) segments as a reference.



(a) Class1(SO) (b) Class2(SO) (c) Class3(SO) (d) Bkg(SO)

(e) Class1(DA) (f) Class2(DA) (g) Class3(DA) (h) Bkg(DA)

Figure 5. TSNE plots of the source-only (SO) variation of our model (top row) and our proposed domain adaptation model (DA) (bottom row). Find in the first 3 columns the TSNE plots of action classes 1 to 3 of the source (red) and target (blue) domain anchors. The last column shows the plot of the background anchors, so those not assigned to any GT label.

5. Conclusions

In this work, we deal for the first time with Unsupervised Domain Adaptation on realistic Temporal Action Localization (TAL) scenarios. We propose a novel semantic adversarial loss that enables a more fine-grained distribution alignment compared to existing global-distribution-based approaches. Given the lack of suitable evaluation setups for this scenario, we propose a suite of 7 different benchmarks that provide a comprehensive assessment of the model performance across various domain shifts. These experiments indicate that our approach yields a considerable improvement over state-of-the-art methods, which we support with extensive quantitative and qualitative results.

Ackwoledgements: This work has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme.

References

[1] Agarwal, N., Chen, Y.T., Dariush, B., Yang, M.H.: Unsupervised domain adaptation for spatio-temporal action localization. arXiv preprint arXiv:2010.09211 (2020) 3

[2] Buch, S., Escorcia, V., Shen, C., Ghanem, B., Carlos Niebles, J.: Sst: Single-stream temporal action proposals. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2911–2920 (2017) 2

[3] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020) 2

[4] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 3

[5] Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1130–1139 (2018) 2

[6] 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. pp. 6321–6330 (2019) 3

[7] Chen, M.H., Li, B., Bao, Y., AlRegib, G., Kira, Z.: Action segmentation with joint self-supervised temporal domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9454–9463 (2020) 2, 3, 5, 7

[8] Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018) 3

[9] Csurka, G.: A comprehensive survey on domain adaptation for visual applications. Domain adaptation in computer vision applications pp. 1–35 (2017) 6

[10] Csurka, G.: Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374 (2017) 2

[11] Cui, S., Wang, S., Zhuo, J., Su, C., Huang, Q., Tian, Q.: Gradually vanishing bridge for adversarial domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12455–12464 (2020) 3

[12] 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: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision (IJCV) **130**, 33–55 (2022), https://doi.org/10.1007/s11263-021-01531-2 2, 5, 6

[13] Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: CVPR 2011. pp. 3281–3288. IEEE (2011) 2, 3, 5

[14] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019) 3

[15] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015) 2, 3, 4, 5

[16] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016) 1, 7

[17] Gao, Z., Zhang, S., Huang, K., Wang, Q., Zhong, C.: Gradient distribution alignment certificates better adversarial domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8937–8946 (2021) 3, 7

[18] Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004) 3

[19] Haeusser, P., Frerix, T., Mordvintsev, A., Cremers, D.: Associative domain adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 2765–2773 (2017) 3

[20] HassanPour Zonoozi, M., Seydi, V.: A survey on adversarial domain adaptation. Neural Processing Letters **55**(3), 2429–2469 (2023) 2, 3

[21] Huang, Z., Wen, J., Chen, S., Zhu, L., Zheng, N.: Discriminative radial domain adaptation. IEEE Transactions on Image Processing **32**, 1419–1431 (2023) 3, 7

[22] Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The thumos challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding **155**, 1–23 (2017) 2

[23] Jamal, A., Namboodiri, V.P., Deodhare, D., Venkatesh, K.: Deep domain adaptation in action space. In: BMVC. vol. 2, p. 5 (2018) 3

[24] Koniusz, P., Tas, Y., Porikli, F.: Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4478–4487 (2017) 3

[25] Kuehne, H., Arslan, A.B., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of Computer Vision and Pattern Recognition Conference (CVPR) (2014) 2, 3, 5

[26] Li, Y., Guo, L., Ge, Y.: Pseudo labels for unsupervised domain adaptation: A review. Electronics **12**(15), 3325 (2023) 2, 3, 4

[27] Li, Z., Yao, L.: Three birds with one stone: Multi-task temporal action detection via recycling temporal annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4751–4760 (2021) 2

[28] Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3320–3329 (2021) 2

[29] Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 988–996 (2017) 2

[30] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 4

[31] Liu, X., Bai, S., Bai, X.: An empirical study of end-to-end temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20010–20019 (2022) 2

[32] Liu, X., Wang, Q., Hu, Y., Tang, X., Bai, S., Bai, X.: End-to-end temporal action detection with transformer. arXiv preprint arXiv:2106.10271 (2021) 2

[33] Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015) 3

[34] Lu, Y., Singh, G., Saha, S., Van Gool, L.: Exploiting instance-based mixed sampling via auxiliary source domain supervision for domain-adaptive action detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4145–4156 (2023) 3

[35] Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5715–5725 (2017) 2, 3

[36] 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) 5, 6

[37] Oza, P., Sindagi, V.A., Sharmini, V.V., Patel, V.M.: Unsupervised domain adaptation of object detectors: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3

[38] Pan, B., Cao, Z., Adeli, E., Niebles, J.C.: Adversarial cross-domain action recognition with co-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11815–11822 (2020) 3

[39] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009) 6

[40] Qing, Z., Su, H., Gan, W., Wang, D., Wu, W., Wang, X., Qiao, Y., Yan, J., Gao, C., Sang, N.: Temporal context aggregation network for temporal action proposal refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 485–494 (2021) 2

[41] Shen, J., Qu, Y., Zhang, W., Yu, Y.: Wasserstein distance guided representation learning for domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 7

[42] Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866 (2023) 2, 3, 4, 6, 7

[43] Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutional-de-convolutional networks

for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5734–5743 (2017) 2

[44] Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018) 2, 6

[45] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 510–526. Springer (2016) 6

[46] Sridhar, D., Quader, N., Muralidharan, S., Li, Y., Dai, P., Lu, J.: Class semantics-based attention for action detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13739–13748 (2021) 2

[47] Tan, J., Tang, J., Wang, L., Wu, G.: Relaxed transformer decoders for direct action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13526–13535 (2021) 2

[48] Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019) 4

[49] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017) 3, 7

[50] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 2

[51] Wei, P., Kong, L., Qu, X., Ren, Y., Jiang, J., Yin, X., et al.: Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 5, 7

[52] Wei, P., Kong, L., Qu, X., Ren, Y., Xu, Z., Jiang, J., Yin, X.: Unsupervised video domain adaptation for action recognition: A disentanglement perspective. Advances in Neural Information Processing Systems **36** (2024) 7

[53] Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: International conference on machine learning. pp. 5423–5432. PMLR (2018) 1, 3, 7

[54] Xu, Y., Cao, H., Mao, K., Chen, Z., Xie, L., Yang, J.: Aligning correlation information for domain adaptation in action recognition. IEEE Transactions on Neural Networks and Learning Systems (2022) 3

[55] Xu, Y., Yang, J., Cao, H., Chen, Z., Li, Q., Mao, K.: Partial video domain adaptation with partial adversarial temporal attentive network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9332–9341 (2021) 3

[56] Yuan, Z., Stroud, J.C., Lu, T., Deng, J.: Temporal action localization by structured maximal sums. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3684–3692 (2017) 2

[57] Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022) 2, 6, 7

[58] Zhang, Y., Doughty, H., Shao, L., Snoek, C.G.: Audio-adaptive activity recognition across video domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13791–13800 (2022) 2, 6

[59] Zhao, H., Torralba, A., Torresani, L., Yan, Z.: Hacs: Human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8668–8678 (2019) 2

[60] Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y., Tian, Q.: Bottom-up temporal action localization with mutual regularization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 539–555. Springer (2020) 2

[61] Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2914–2923 (2017) 2