

AutoLabel: CLIP-based framework for Open-set Video Domain Adaptation

Giacomo Zara¹, Subhankar Roy³, Paolo Rota¹, Elisa Ricci^{1,2}

¹University of Trento, Italy ²Fondazione Bruno Kessler, Italy

³LTCI, Télécom Paris, Institut polytechnique de Paris, France

{giacomo.zara,paolo.rota,e.ricci}@unitn.it, subhankar.roy@telecom-paris.fr

Abstract

Open-set Unsupervised Video Domain Adaptation (OUVDA) deals with the task of adapting an action recognition model from a labelled source domain to an unlabelled target domain that contains “target-private” categories, which are present in the target but absent in the source. In this work we deviate from the prior work of training a specialized open-set classifier or weighted adversarial learning by proposing to use pre-trained Language and Vision Models (CLIP). The CLIP is well suited for OUVDA due to its rich representation and the zero-shot recognition capabilities. However, rejecting target-private instances with the CLIP’s zero-shot protocol requires oracle knowledge about the target-private label names. To circumvent the impossibility of the knowledge of label names, we propose AutoLabel that automatically discovers and generates object-centric compositional candidate target-private class names. Despite its simplicity, we show that CLIP when equipped with AutoLabel can satisfactorily reject the target-private instances, thereby facilitating better alignment between the shared classes of the two domains. The code is available¹.

1. Introduction

Recognizing actions in video sequences is an important task in the field of computer vision, which finds a wide range of applications in human-robot interaction, sports, surveillance, and anomalous event detection, among others. Due to its high importance in numerous practical applications, action recognition has been heavily addressed using deep learning techniques [32, 44, 58]. Much of the success in action recognition have noticeably been achieved in the supervised learning regime [6, 17, 49], and more recently shown to be promising in the unsupervised regime [20, 34, 56] as well. As constructing large scale annotated and curated action recognition datasets is both challenging and expensive, focus has shifted towards adapting a model from a source domain, having a labelled source dataset, to an un-

labelled target domain of interest. However, due to the discrepancy (or *domain shift*) between the source and target domains, naive usage of a source trained model in the target domain leads to sub-optimal performance [51].

To counter the domain shift and improve the transfer of knowledge from a labelled source dataset to an unlabelled target dataset, unsupervised video domain adaptation (UVDA) methods [7, 9, 42] have been proposed in the literature. Most of the prior literature in UVDA are designed with the assumption that the label space in the source and target domain are identical. This is a very strict assumption, which can easily become void in practice, as the target domain may contain samples from action categories that are not present in the source dataset [43]. In order to make UVDA methods more useful for practical settings, open-set unsupervised video domain adaptation (OUVDA) methods have recently been proposed [5, 8]. The main task in OUVDA comprise in promoting the adaptation between the *shared* (or *known*) classes of the two domains by excluding the action categories that are exclusive to the target domain, also called as *target-private* (or *unknown*) classes. Existing OUVDA prior arts either train a specialized open-set classifier [5] or weighted adversarial learning strategy [8] to exclude the target-private classes.

Contrarily, we address OUVDA by tapping into the very rich representations of the open-sourced foundation *Language and Vision Models* (LVMs). In particular, we use CLIP (Contrastive Language-Image Pre-training) [45], a foundation model that is trained on *web-scale* image-text pairs, as the core element of our framework. We argue that the LVMs (*e.g.*, CLIP) *naturally* lend themselves well to OUVDA setting due to: (i) the representation learned by LVMs from *webly* supervised *image-caption* pairs comes encoded with an immense amount of prior about the real-world, which is (un)surprisingly beneficial in narrowing the shift in data distributions, even for video data; (ii) the zero-shot recognition capability of such models facilitates identification and separation of the target-private classes from the shared ones, which in turn ensures better alignment between the known classes of the two domains.

¹<https://github.com/gzaraunitn/autolabel>

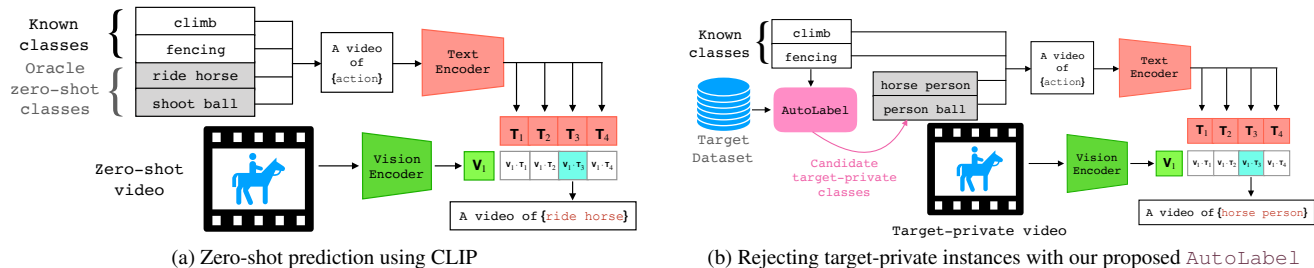


Figure 1. Comparison of `AutoLabel` with CLIP [45] for zero-shot prediction on target-private instances. (a) CLIP assumes the knowledge about the *oracle* zero-shot classes names (ride horse, shoot ball); (b) Our proposed `AutoLabel` discovers automatically the *candidate* target-private classes (horse person, person ball) and extends the known classes label set

Zero-shot inference using CLIP requires multi-modal inputs, *i.e.*, a test video and a set of *all* possible prompts “A video of {label}”, where *label* is a class name, for computing the cosine similarity (see Fig. 1a). However in the OUVDA scenario, except the shared classes, *a priori* knowledge about the target-private classes label names are not available (the target dataset being unlabelled). Thus, exploiting zero-shot capability of CLIP to identify the target-private instances in an unconstrained OUVDA scenario becomes a bottleneck. To overcome this issue we propose `AutoLabel`, an automatic labelling framework that constructs a set of *candidate* target-private class names, which are then used by CLIP to potentially identify the target-private instances in a zero-shot manner.

In details, the goal of `AutoLabel` is to augment the set of shared class names (available from the source dataset) with a set of candidate target-private class names that best represent the true target-private class names in the target dataset at hand (see Fig. 1b). To this end, we use an external pre-trained *image captioning* model ViLT [31] to extract a set of *attribute* names from every frame in a video sequence (see Sec. 3.2.1 for details). This is motivated by the fact that actions are often described by the constituent objects and actors in a video sequence. As an example, a video with the prompt “A video of {chopping onion}” can be *loosely* described by the proxy prompt “A video of {knife}, {onion} and {arm}” crafted from the predicted attribute names. In other words, the attributes “knife”, “onion” and “arm” when presented to CLIP in a prompt can elicit similar response as the true action label “chopping onion”.

Naively expanding the label set using ViLT predicted attributes can introduce redundancy because: (i) ViLT predicts attributes per frame and thus, there can be a lot of distractor object attributes in a video sequence; and (ii) ViLT predicted attributes for the shared target instances will be duplicates of the true source action labels. Redundancy in the shared class names will lead to ambiguity in target-private instance rejection.

Our proposed framework `AutoLabel` reduces the redundancy in the effective label set in the following manner.

First, it uses unsupervised clustering (*e.g.*, k-means [36]) on the target dataset to cluster the target samples, and then constructs the top-*k* most frequently occurring attributes among the target samples that are assigned to each cluster. This step gets rid of the long-tailed set of attributes, which are inconsequential for predicting an action (see 3.2.2 for details). Second, `AutoLabel` removes the duplicate sets of attributes that bear resemblance with the source class names (being the same shared underlying class) by using a set *matching* technique. At the end of this step, the effective label set comprises the shared class names and the candidate sets of attribute names that represent the target-private class names (see Sec. 3.2.3 for details). Thus, `AutoLabel` unlocks the zero-shot potential of the CLIP, which is very beneficial in unconstrained OUVDA.

Finally, to transfer knowledge from the source to the target dataset, we adopt conditional alignment using a simple *pseudo-labelling* mechanism. In details, we provide to the CLIP-based encoder the target samples and the extended label set containing the shared and candidate target-private classes. Then we take the top-*k* pseudo-labelled samples for each predicted class and use them for optimizing a supervised loss (see Sec. 3.2.4 for details). Unlike many open-set methods [5, 8] that reject all the target-private into a single unknown category, `AutoLabel` allows us to discriminate even among the target-private classes. Thus, the novelty of our `AutoLabel` lies not only in facilitating the rejection of target-private classes from the shared ones, but also opens doors to open world recognition [1].

In summary, our contributions are: (i) We demonstrate that the LVMS like CLIP can be harnessed to address OUVDA, which can be excellent replacement to complicated alignment strategies; (ii) We propose `AutoLabel`, an automatic labelling framework that discovers candidate target-private classes names in order to promote better separation of shared and target-private instances; and (iii) We conduct thorough experimental evaluation on multiple benchmarks and surpass the existing OUVDA state-of-the-art methods.

2. Related Work

Action Recognition. A plethora of deep learning methods have been proposed for action recognition, which can

be roughly categorized based upon the type of network architecture used to process the video clips and the input modalities (see [32] for an extensive survey). The most common methods rely on 2D-CNNs [29, 54] coupled with frames aggregation methods [14, 23, 28], and on 3D-CNNs [6, 24, 27, 52]. Furthermore, the task has been addressed with two-stream methods introducing optical flow [18, 49] and more recently with transformer-based architectures [4, 22, 40]. Despite the impressive success, these models rely on large annotated datasets to train, which is indeed a bottleneck when no annotations are available. Our work focuses on adapting a model to an unlabelled target domain by exploiting the knowledge from a labelled source domain.

Open-set Unsupervised Domain Adaptation. Mitigating domain-shift with unsupervised domain adaptation methods has been an active area of research for both images (see survey in [10]) and videos [5, 7–9, 12, 42]. In particular, closed-set video domain adaptation has been realized using the pretext task of clip order prediction [9], adversarial learning coupled with temporal attention [7, 41] and contrastive learning with multi-stream networks [30, 42], among others. Closed-set domain adaptation being unrealistic, adaptation methods have been proposed to address the open-set adaptation scenario, but are mostly limited to the image classification task (see survey in [11]). However, for OUVDA the prior work [5, 8] is quite scarce. Busto *et al.* [5] proposed a method that first learns a mapping from the source to the target. After the transformation is learned, linear one-vs-one SVMs are used to separate the unknown classes. CEVT [8] tackles OUVDA by modelling the entropy of the target samples as generalised extreme value distribution, with the target-private samples lying at the tail of the distribution. The entropy is then used as weights to conduct weighted adversarial alignment. Differently, our work leans on the excellent zero-shot capability of CLIP [45] to detect the target-private instances. Post detection, we resort to conditional alignment to align the shared classes in the source and target domains using pseudo-labelling.

Language and Vision Models. Off late, the vision community has witnessed a paradigm shift with the advent of the language and vision (LVM) foundation models [2]. LVMs derive strength from the large scale pre-training with web-scale multi-modal image-text or image-audio training pairs. In particular, CLIP [45] has demonstrated excellent downstream performance on zero-shot classification. Inspired by its success, CLIP has been adapted as ActionCLIP [55] and VideoCLIP [57] for addressing zero-shot action recognition and action segmentation, respectively. While ActionCLIP assumes knowledge about the oracle zero-shot classes, VideoCLIP uses a thresholded closed-set predictor to identify the unknown classes. We argue that both are impractical and sub-optimal, and therefore propose to discover

the target-private classes in an automatic manner. Very recently, Esmailpour *et al.* [16] indeed proposed to detect zero-shot classes without any oracle knowledge, but limit themselves to image classification task only. In contrast, our `AutoLabel` takes additional care in reducing redundancy and is thought-out for the action recognition task.

3. Methods

In this work we propose `AutoLabel` to tackle the task of adapting a model from a labelled source dataset to an unlabelled target dataset, under the constraint that the target contains samples from action categories that are not present in the source domain. Before we describe the details of `AutoLabel` we formalize OUVDA and present the preliminaries used in our framework.

Problem Definition and Notations. Let us assume that we are given a source dataset containing *labelled* video sequences $\mathcal{D}^S = \{(\mathbf{X}_i^S, l_i^S)\}_{i=1}^n$, where $\mathbf{X} \in \mathcal{X}$ represents the input video and $l^S \in \mathcal{Y}^S = \{l_1, l_2, \dots, l_K\}$ being the K *shared* class names, instead of class indices. For *e.g.*, $\mathcal{Y}^S = \{\text{“climb”}, \text{“fencing”}, \dots, \text{“push up”}\}$. Moreover, we are also given an *unlabelled* target dataset $\mathcal{D}^T = \{\mathbf{X}_i^T\}_{i=1}^n$ containing n samples from classes \mathcal{Y}^T . In particular, $\mathcal{Y}^T = \mathcal{Y}^S \cup \mathcal{Y}^U$, where $\mathcal{Y}^U = \{l_{K+1}, l_{K+1}, \dots, l_{K+M}\}$ represents the M *target-private* class names and are not known to us *a priori*. Each video sequence is composed of r frames $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^r$ depicting an action with label l .

The goal in OUVDA is to learn a parameterized function $f_\theta: \mathcal{X} \rightarrow \mathcal{Y}$ using $\mathcal{D}^S \cup \mathcal{D}^T$, that can correctly predict the shared target instances to be belonging to one of the classes in \mathcal{Y}^S and reject the target-private instances as “unknown”.

Overview. To address OUVDA we propose `AutoLabel` (see Fig. 2), a CLIP-based framework, that comprise of a transformer [53] as the text encoder and ViT [15] as the vision encoder. In practice, we use the ActionCLIP architecture [55] that has an additional self-attention temporal pooling to aggregate the frame-level features to output video-level feature. To enable the identification of target-private instances without any access to \mathcal{Y}^U , `AutoLabel` generates a set of candidate target-private classes names.

In details, `AutoLabel` uses pre-trained ViLT [31] to predict a set of attributes from the frames in target video sequences. Then it uses unsupervised clustering on the video-level target features to cluster the video sequences into respective semantic clusters. All the predicted frame-level attributes, pertaining to the video sequences that are assigned to a given cluster, are filtered to yield the top- k most salient attributes. These top- k attributes are then concatenated to form a *proxy* action label name. Note that this step will produce one candidate label name per cluster, including the ones corresponding to the shared classes. To disambiguate and merge the redundant class label names with the known shared class names we use set matching. At the end of this

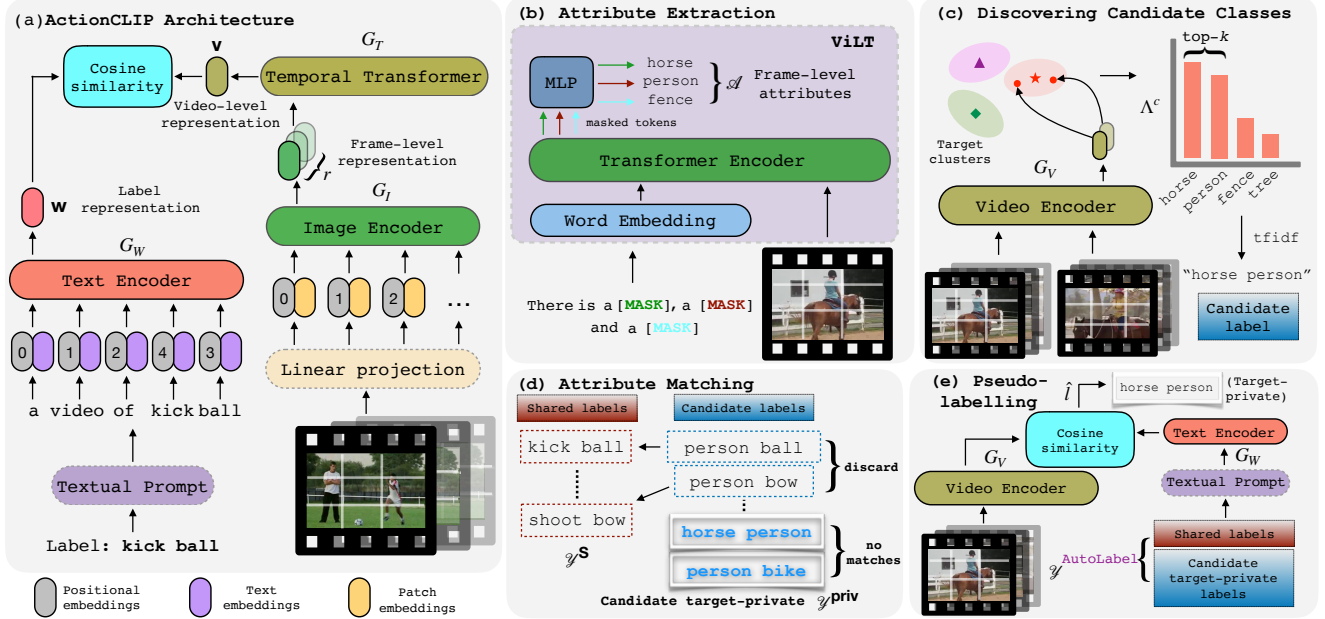


Figure 2. Overview of the `AutoLabel` framework for OUVDA: (a) the ActionCLIP [55] forms the backbone of `AutoLabel` for extracting video-level and text label representations; (b) per-frame attributes are extracted using ViLT [31]; (c) candidate class names in the target dataset are discovered in an unsupervised manner; (d) redundant candidate labels are discarded and candidate target-private class names are identified with attribute matching; and (e) extended label set $\mathcal{Y}^{\text{AutoLabel}}$ enables the rejection of target-private instances

step the effective label set will comprise of the shared class names and the candidate target-private class names.

Following the extended label set creation, pseudo-labels are computed for the unlabelled target samples by providing to the text and vision encoders of ActionCLIP, as input, the target video sequences and the extended label set, respectively. The top- $k\%$ most confident predictions from each predicted class (computed using argmax on the output probability distribution) are then used as hard pseudo-labels for optimizing the ActionCLIP multi-modal training objective. Below we summarize the ActionCLIP.

3.1. Preliminaries

ActionCLIP [55] is a transformer-based multi-modal supervised action recognition framework that consist of a text encoder $G_W(\cdot)$ to extract features of text labels and a video encoder $G_V(\cdot)$ to extract spatio-temporal features from a corresponding video (see Fig. 2a). The video encoder consists of an image encoder $G_I(\cdot)$, operating at frame-level, and a temporal transformer $G_T(\cdot)$ for aggregation, such that $G_V = G_T \circ G_I$. ActionCLIP is trained to maximize the cosine similarity $\text{cos}(\cdot, \cdot)$ between the pairwise video and label representations, which is defined as:

$$\text{cos}(\mathbf{X}, l) = \frac{\mathbf{v} \cdot \mathbf{w}^T}{\|\mathbf{v}\| \|\mathbf{w}\|} \quad (1)$$

where $\mathbf{v} = G_V(\mathbf{X})$ and $\mathbf{w} = G_W(l)$ are the feature representations of the video and text modalities, respectively. The softmax-normalized similarity score for the i^{th} training pair is given as:

$$p_i(\mathbf{X}_i, l_i) = \frac{\exp(\text{cos}(\mathbf{X}_i, l_i)/\tau)}{\sum_{j=1}^N \exp(\text{cos}(\mathbf{X}_i, l_j)/\tau)} \quad (2)$$

where N is the number of training pairs in a mini-batch and τ is a temperature hyperparameter. The ground-truth similarity score q is 1 and 0 for positive and negative pairs, respectively. The symmetric video-text contrastive loss used to train ActionCLIP on the source dataset is defined with the Kullback-Leibler (KL) divergence as:

$$\mathcal{L}_{\text{ActionCLIP}} = \frac{1}{2} \mathbb{E}_{(\mathbf{X}^s, l^s) \sim \mathcal{D}^s} [\text{KL}(\mathbf{p}|\mathbf{q}) + \text{KL}(\mathbf{q}|\mathbf{p})] \quad (3)$$

3.2. AutoLabel Framework

The main task of the proposed `AutoLabel` framework is to equip the ActionCLIP with zero-shot detection capabilities such that the resulting framework can be used in OUVDA, without any oracle knowledge about the target-private class names \mathcal{Y}^u . To this end, `AutoLabel` relies on a series of four sub-modules: (i) Attribute extraction is in charge of extracting a set of frame-level attributes (e.g., actors and objects) depicted in a video frame (see Fig. 2b); (ii) the second sub-module (see Fig. 2c) uses the predicted attributes to construct a set of candidate labels for the target domain that correspond to different semantic categories; (iii) Attribute matching sub-module (see Fig. 2d) further reduces the redundancy in the candidate label names that are duplicates of the known shared class names \mathcal{Y}^s ; and (iv) Pseudo-labelling (see Fig. 2e) sub-module then uses

the shared and candidate target-private class names to reject the target-private instances. Next we elaborate each sub-module of `AutoLabel` in detail.

3.2.1 Attribute Extraction

As discussed in Sec. 1, our automatic labelling approach is motivated by the fact that actions are often times described by the *objects* and *actors* in a video. Thus, modelling the attributes alone and constructing the candidate label names from such attributes can suffice to reject target-private instances, which is one of the two main goals in OUVDA. We use an off-the-shelf image captioning model ViLT [31] and prompt it in a way to obtain a set of attributes per frame.

Specifically, we give to the ViLT model a video frame \mathbf{x}_j^T and a prompt $z = \text{“There is a [MASK], a [MASK] and a [MASK].”}$. The model outputs a set of most probable words for the m masked tokens as:

$$\mathcal{A}(\mathbf{x}_j^T) = \text{ViLT}(\mathbf{x}_j^T, z) \quad (4)$$

where \mathcal{A} denote a set of attributes with $m = \text{card}(\mathcal{A})$. As an example in the Fig. 2b, the ViLT predicts $m = 3$ attributes: “horse”, “person” and “fence” corresponding to the three masked tokens in the prompt z .

In a similar spirit, ZOC [16] used an image captioning model to generate attributes for enabling zero-shot prediction. However, we differ from ZOC in the following ways: (i) ZOC addresses image classification whereas we tackle action recognition, (ii) ZOC treats each attribute as a candidate class, while we create compositional candidate class names by combining multiple attributes. This is crucial in action recognition because action names arise from the interaction of objects and actors; and (iii) Unlike us, ZOC does not handle the redundancy caused by duplicate candidate label and shared class names. Next we describe how `AutoLabel` combines attributes to create candidate action label names and how such redundancies can be eliminated.

3.2.2 Discovering Candidate Classes

As mentioned above, an action label name is a product of the interaction between object(s) and actor(s). For instance, the action label “ride horse” depicts an interaction between a “horse” and a “person”. If the attributes in \mathcal{A} are treated as candidate labels in isolation, like ZOC, then for a video of horse riding the cosine similarity between the visual and text representations of both “horse” and “person” text will be high. Moreover, there can be other distractor attributes such as “fence”, which if present in a frame, will also elicit high response from CLIP. In order to uniquely model the true label “ride horse” we propose the following strategy.

First, we use the video encoder G_V to cluster all the target videos of \mathcal{D}^T into \mathcal{C} target clusters, which ideally should represent semantic categories. Note that we do not assume *a priori* knowledge about the true cardinality of $|\mathcal{Y}^U|$ and

we set $|\mathcal{C}| > |\mathcal{Y}^S|$. In details, we use the standard clustering algorithm k -means [37] that takes as input a set of the video-level features $\mathbf{v}^T = G_V(\mathbf{X}^T)$ and assigns them to \mathcal{C} distinct centroids μ_c , with the cluster assignment for the i^{th} video sequence given as $y_i \in \{0, 1\}^{|\mathcal{C}|}$. Next, we construct a histogram per target cluster (see Fig. 2c), by using all the attributes $\Lambda^{c,T} = \{\mathcal{A}(\mathbf{x}^T) | \mathbf{x}^T \in \mathbf{X}_{\hat{y}^{c,T}}^T, \mathbf{X}^T \in \mathcal{D}^T\}$ associated to a target cluster c , where $\hat{y}^{c,T} = \arg \min_{c \in \mathcal{C}} \|\mu_c - \mathbf{X}^T\|$. Note that this step is carried out at the beginning of each training epoch.

We expect the most frequent attributes associated to a cluster to be the most salient and descriptive of the action. As shown in Fig. 2c, in the “ride horse” cluster, “horse” and “person” will be the most frequent attributes, and the rest will lie in the tail of the distribution. We filter the t most common and relevant attributes in $\Lambda^{c,T}$ to obtain $\bar{\Lambda}^{c,T} = \text{tfidf}(\text{argtop}_k(\Lambda^{c,T}))$. Finally, we concatenate the attributes in $\bar{\Lambda}^{c,T}$ to form the candidate label:

$$l_c^{\text{cand},T} = \bar{\Lambda}_1^{c,T} || \dots || \bar{\Lambda}_t^{c,T} \quad (5)$$

where $|| \cdot ||$ represent the concatenation operation separated by a space. Details about the `tfidf`(\cdot) operator is provided in the **Supp Mat**. Since, the target is unlabelled, we can not yet distinguish the shared candidate labels from the target-private ones. Thus, to identify the target-private instances we need a mechanism to tell apart the target-private class names, which we describe next.

3.2.3 Attribute Matching

The attribute matching step is in charge of finding the candidate label names that correspond to the target-private classes. To this end, we simply find the candidate label names in $\mathcal{Y}^{\text{cand},T} = \{l_1^{\text{cand},T}, \dots, l_{|\mathcal{C}|}^{\text{cand},T}\}$ that correspond to the shared label names in the source \mathcal{Y}^S . This will essentially leave us with the candidate target-private label names that have no match with the source (see Fig. 2d).

In details, we repeat the sub-module described in Sec. 3.2.2 on the source samples to obtain the set of attributes $\bar{\Lambda}^{l^S}$, where l^S is a source label name. Then we create a similarity matrix $S \in \mathbb{R}^{K \times |\mathcal{C}|}$, where an entry $s_{i,j}$ in S denotes how similar a i^{th} source label name l_i^S is to a j^{th} candidate label name $l_j^{\text{cand},T}$, and is formally defined as:

$$s_{i,j} = \text{sim}(\underbrace{\bar{\Lambda}^{l_i^S}}_{\text{attributes of } i^{\text{th}} \text{ source label}}, \underbrace{\bar{\Lambda}^{l_j^{\text{cand},T}}}_{\text{attributes of } j^{\text{th}} \text{ target label}}) \quad (6)$$

where $\text{sim}(\cdot, \cdot)$ is a scoring function that computes a similarity score based on the common attributes between the two sets and their occurrences. More details about the $\text{sim}(\cdot, \cdot)$ can be found in the **Supp Mat**. If the score is higher than a threshold γ , the two sets are considered as matched, and the j^{th} target candidate label is discarded. After this thresholding step, the candidate labels in $\mathcal{Y}^{\text{cand},T}$ that are not matched

with any of the source labels become the candidate target-private label names, such that $\mathcal{Y}^{\text{priv},T} \subset \mathcal{Y}^{\text{cand},T}$.

3.2.4 Conditional Alignment

After the attribute matching, we have an extended label set $\mathcal{Y}^{\text{AutoLabel}} = \mathcal{Y}^S \cup \mathcal{Y}^{\text{priv},T}$, which comprise of the shared labels \mathcal{Y}^S and the candidate target-private labels $\mathcal{Y}^{\text{priv},T}$. The CLIP is now equipped to detect target-private instances. Specifically, for a target sample \mathbf{X}^T we compute the predicted label (see Fig. 2e) as:

$$\hat{l} = \arg \max_{l \in \mathcal{Y}^{\text{AutoLabel}}} \frac{\exp(\cos(\mathbf{X}^T, l)/\tau)}{\sum_{j=1}^{|\mathcal{Y}^{\text{AutoLabel}}|} \exp(\cos(\mathbf{X}^T, l_j)/\tau)} \quad (7)$$

where the \mathbf{X}^T is considered as a target-private if $\hat{l} \in \mathcal{Y}^{\text{priv},T}$.

To align the shared classes of the two domains we resort to a simple conditional alignment strategy with the usage of pseudo-labels computed using Eq. (7). In practice, we extract the top- $k\%$ most confident pseudo-labels per predicted class and backpropagate the supervised ActionCLIP loss in Eq. (3) for the pseudo-labelled target samples. Besides the shared classes, this loss also promotes discrimination among the target-private classes as gradients are also backpropagated corresponding to the discovered candidate target-private classes $\mathcal{Y}^{\text{priv},T}$. This is indeed promising from the open world recognition point of view because the target-private samples are assigned to their respective semantic categories, instead of a single “unknown” category.

4. Experiments

Datasets and Settings. We evaluate our proposed method and the baselines on two benchmarks derived from the datasets HMDB [33], UCF101 [50] and Epic-Kitchens [13], for the OUVDA task. The first OUVDA benchmark is the *HMDB↔UCF*, introduced by Chen *et al.* [8], that comprises action categories, where six of them are shared between the two domains and the remaining six are target-private. The second benchmark is the *Epic-Kitchens* (EK) that is composed of egocentric videos from three kitchen environments (or domains D1, D2 and D3). We extend the closed-set UVDA EK benchmark, used in [30, 42], to the OUVDA scenario where there are eight shared action categories and all the remaining ones in each kitchen environment are considered as target-private. The statistics of the two benchmarks are provided in the **Supp Mat**.

Implementation details. As discussed in the Sec. 3.1, we employ the network architecture from ActionCLIP [55], which is composed of transformer-based video and text encoders. In particular, the video encoder contains a ViT-B/32 vision transformer, which is a pre-trained CLIP encoder [45], while the temporal transformer is initialized from scratch. The label representations are obtained with a 512-wide transformer containing 8 attention heads.

We extract five attributes ($m = 5$) from the pre-trained image captioning model ViLT [31], and set the number of attributes t in a candidate label to 5 and 2 for *HMDB↔UCF* and EK, respectively. To train our framework, we used AdamW optimizer [39], with a learning rate of 5×10^{-5} and a weight decay rate of 0.2. The models were trained for 20 epochs using a total mini-batch size of 96. Each training video sequence is composed of randomly sampled $r = 8$ frames, each of resolution 224×224 .

Evaluation Metrics. We evaluate the performance of the models on the target dataset using standard open-set metrics, as in [8]. In particular, we report: the **ALL** accuracy which is the percentage of correctly predicted target samples over all the target samples; the closed-set **OS*** accuracy which computes the averaged accuracy over the known classes only; the **UNK** recall metric which denotes a ratio of the number of correctly predicted unknown samples over the total number of unknown samples in the target dataset; and **HOS** is the harmonic mean between **OS*** and **UNK** metrics, *i.e.*, $\text{HOS} = 2 \times \frac{\text{OS}^* \times \text{UNK}}{\text{OS}^* + \text{UNK}}$. As the **HOS** takes into account both the closed and open-set scores, it is considered as the most meaningful metric for evaluating open-set algorithms [3, 8, 46].

4.1. Comparison with the State of the Art

Baselines. We compare our *AutoLabel* with CEVT [8], an existing state-of-the-art method for OUVDA. However, note that CEVT uses ResNet-101 [25] (pre-trained on ImageNet-1k) as a backbone for extracting frame-level features, which is weaker w.r.t. the CLIP pre-trained backbone used by *AutoLabel*. To be fairly comparable, we create a baseline **CEVT-CLIP** that replaces the ResNet-101 backbone of CEVT with the stronger CLIP vision encoder.

Additionally, we introduce few more baselines for OUVDA that use the representation power of CLIP, but *without* the *a priori* knowledge about the true target-private label set names. These baselines differ by how the target-private instances are rejected: (i) the **ActionCLIP** baseline that rejects target-private instances by thresholding the similarity values computed using the shared class names, (ii) the **ActionCLIP-ZOC** baseline implements the target-private rejection mechanism of ZOC [16]; and (iii) the **ActionCLIP-Oracle** that assumes the knowledge of the true target-private label set names, as described in [19]. Note that all these baselines then fine-tune on the pseudo-labelled target data, similar to our proposed *AutoLabel* framework.

For the *HMDB↔UCF* adaptation scenario we also include the baselines reported in [8], which are closed-set [7, 21, 38] and open-set methods [48] adapted to the OUVDA setting, accordingly.

Results on *HMDB↔UCF*. In Tab. 1 we report the results obtained on the *HMDB↔UCF* benchmark. We separately report the baselines that use ResNet101 and CLIP back-

Method	Backbone	$HMDB \rightarrow UCF$				$UCF \rightarrow HMDB$			
		ALL	OS*	UNK	HOS	ALL	OS*	UNK	HOS
DANN [21] + OSVM [26]	ResNet101 [25]	64.6	62.9	74.7	68.3	66.1	48.3	83.9	61.3
JAN [38] + OSVM [26]		61.5	62.9	73.8	67.9	61.1	47.8	74.4	58.2
AdaBN [35] + OSVM [26]		62.9	58.8	73.3	65.3	62.9	58.8	73.3	65.3
MCD [47] + OSVM [26]		66.7	63.5	73.8	68.3	66.7	57.8	75.6	65.5
TA ² N [7] + OSVM [26]		63.4	61.3	79.0	69.1	65.3	56.1	74.4	64.0
TA ³ N [7] + OSVM [26]		60.6	58.4	82.5	68.4	62.2	53.3	71.7	61.2
OSBP [48] + AvgPool		64.8	55.3	85.7	67.2	67.2	50.8	84.5	63.5
CEVT [8]		70.6	66.8	84.3	74.5	75.3	56.1	94.5	70.4
CEVT-CLIP [8]	CLIP [45]	70.9	68.0	92.5	78.4	78.3	61.2	92.1	73.5
ActionCLIP [55]		79.7	81.1	94.3	87.2	84.8	75.7	91.5	82.9
ActionCLIP-ZOC [16]		74.7	84.0	68.5	75.5	85.4	75.5	93.2	83.4
AutoLabel (ours)		79.7	82.5	94.3	88.0	86.0	82.9	88.2	85.5
ActionCLIP-Oracle [19]	CLIP [45]	93.3	93.7	100.0	96.7	92.8	86.5	98.3	92.0

Table 1. Comparison with the state-of-the-art on the $HMDB \leftrightarrow UCF$ benchmark for OUVDA. The best performances of the CLIP-based and ResNet-based methods are shown in bold and underlines, respectively. The *Oracle* performance is shown in italics. Using CLIP as backbone greatly improves the HOS scores. The target-private rejection with our `AutoLabel` outperforms all the baseline methods

Method	<i>Epic-Kitchens</i>						Avg
	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	
CEVT [8]	13.2	14.7	8.4	16.0	8.1	11.3	12.0
CEVT-CLIP [8]	13.2	17.4	13.3	14.3	10.2	10.1	13.0
ActionCLIP [55]	31.3	28.3	38.1	43.4	29.0	24.2	32.4
ActionCLIP-ZOC [16]	25.9	26.8	31.7	41.3	28.0	32.1	30.9
AutoLabel (ours)	34.8	38.3	44.1	50.4	31.9	29.8	38.2
ActionCLIP-Oracle [19]	33.2	33.1	37.1	44.6	24.2	28.9	33.5

Table 2. Comparison with the state-of-the-art on the *Epic-Kitchens* benchmark for OUVDA. The HOS scores are reported for all the methods. The best performances are shown in bold. Overall, our `AutoLabel` surpasses all the baselines, including the *Oracle*

bones as feature extractors. It is evident that the usage of CLIP-based backbone can significantly improve the HOS scores over the baselines that use ResNet101 as feature extractor. For *e.g.*, by changing the backbone alone, the **CEVT-CLIP** is +3.9 and +3.1 points better than the **CEVT** for the settings $HMDB \rightarrow UCF$ and $UCF \rightarrow HMDB$, respectively. This highlights the importance of the LVMs in closing the domain gap for the OUVDA task.

Next, we demonstrate that OUVDA can further benefit from CLIP by successfully rejecting the target-private instances with our `AutoLabel`, where we observe +9.6 and +12.0 points improvement over **CEVT-CLIP** on the two settings, as far as the HOS scores are concerned. Our `AutoLabel` also surpasses the **ActionCLIP** baseline in HOS score (by +0.8 and +2.6 points), which does not model the unknown classes and uses a threshold-based target-private rejection with the closed-set classifier. Interestingly, the **ActionCLIP-ZOC** baseline, which indeed models the unknown classes using externally trained image-captioning model like `AutoLabel`, exhibits inferior performance than `AutoLabel` on both the occasions (-12.5 and -2.1 points). This emphasizes the importance of carefully modelling the candidate target-private label names for the action recognition task, which is lacking in ZOC [16], as previously discussed in Sec. 3.2.1.

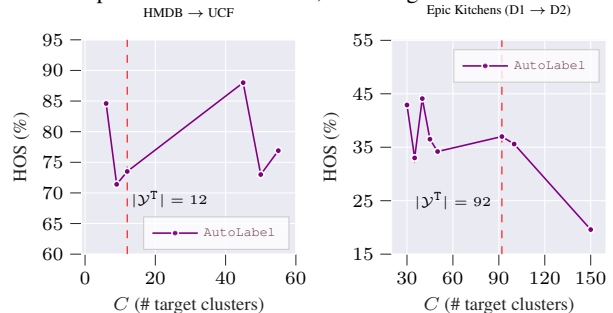


Figure 3. Impact of varying the number of target clusters C .

Finally, when the oracle target-private label set names are available, CLIP demonstrates even stronger performance, as shown in the last row of Tab. 1. However, when such privileged information about target datasets is not available, our `AutoLabel` can effectively close the gap, reaching closer to the upper bound performance.

Results on *Epic-Kitchens*. In Tab. 2 we report the results obtained on the more challenging *Epic-Kitchens* (EK) benchmark. Due to the lack of space we only report the HOS scores for the six directions of EK. Detailed metrics can be found in the **Supp Mat**. The observations on the EK benchmark is consistent with the trends observed in the $HMDB \leftrightarrow UCF$ benchmark. In particular, the **ActionCLIP** baseline achieves a much superior per-

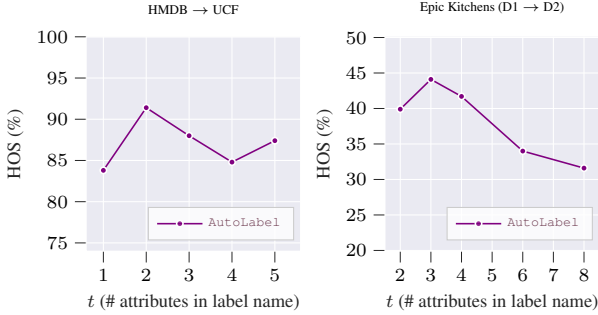


Figure 4. Impact of varying the **number of attributes t in candidate label name**.

formance (+20.4 points) over the existing prior art **CEVT**. Moreover, when ActionCLIP is powered by our proposed **AutoLabel**, the average HOS score further improves from 32.4 to 38.2. Note that the low overall performance on EK by every method hints at the fact that the actions depicted by the egocentric videos in EK are far more challenging. Despite the complexity in such actions, our proposed **AutoLabel** sets a new state-of-the-art performance in OUVDA, surpassing the competitors by a large margin.

Contrary to our previous observation, the **ActionCLIP-Oracle** baseline is outperformed by our **AutoLabel**. We hypothesize that the true target-private label names in the EK (e.g., *flatten*) are quite generic and not descriptive enough to reject the target-private instances. Whereas, our **AutoLabel** models the objects associated with such actions (e.g., “rolling pin” for *flatten*), leading to an improved disambiguation of such action categories.

4.2. Ablation Analysis

Number of target clusters. Given the significant role of the clustering step for semantically modeling the target domain, we show in Fig. 3 the impact of varying the number of target clusters \mathcal{C} on the HOS scores for the *HMDB*→*UCF* and EK (D1 → D2) settings. The dashed red line indicates the ground truth number of target categories, i.e., $|\mathcal{C}| = |\mathcal{Y}^T|$. We can observe that, in both cases, the HOS peaks for values of $|\mathcal{C}|$ in the range 35-45, which are higher and lower than the actual number of target classes for *HMDB*→*UCF* and *Epic-Kitchens*, respectively.

For *HMDB*→*UCF* these findings suggest that due to the wide diversity of actions, the discrimination of the framework is increased by modeling the target domain at a higher granularity. Whereas, for the EK, due to visual similarity of multiple actions, the discovered granularity in the label set is far less excessive than the ground truth action categories.

Number of tokens in the candidate target-private labels.

In Fig. 4 we ablate on the number of attributes t used to build the final target-private label candidate. In line with our hypothesis in Sec. 3.2.1, the ablation suggests that the candidate target labels composed by 2-3 attribute tokens to be

a reasonable choice for effectively describing the candidate target action label names. In particular, when the value of t increases significantly, the HOS scores for both the benchmarks show a steady decline due to the overcharacterization of the action labels with spurious attributes.

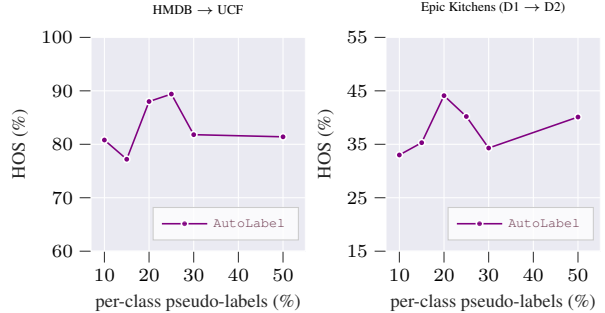


Figure 5. Impact of varying the **number of pseudo-labels**.

Number of target pseudo-labels. Finally, in Fig. 5 we analyse the impact of the number of most confident pseudo-labels employed in order to fine-tune the model on the unlabelled target domain. In both the considered settings, we notice that the optimal percentage of pseudo-labels to be used for target fine-tuning lies around 20%. Further addition of pseudo-labels leads to reduced performance, possibly due to the accumulation of noisy target pseudo-labels.

Limitations

The proposed **AutoLabel** models the target-private label name by indirectly modelling the objects and actors depicted in an action. Currently it can not disambiguate among different actions that involve the same objects and actors. For instance, “polo”, “equestrian” and “skijoring” will all be described by the same candidate target-private label “horse person”. Exploring the relationships between the actors and objects, along with their states, remains as a future work to overcome this limitation.

5. Conclusions

In this work we presented **AutoLabel**, a CLIP-based automatic labelling framework for addressing the open-set unsupervised video domain adaptation. Our proposed **AutoLabel** automatically discovers the candidate target-private class label names and extends the shared class names in order to equip CLIP with zero-shot prediction capabilities, which is indeed necessary for rejecting target-private instances. Empirically we show that **AutoLabel** enabled CLIP models bode well for the OUVDA task due to their rich representations and zero-shot detection capabilities.

Acknowledgment. This work was supported by the AI@TN project, the EU H2020 project SPRING funded by the European Commission under GA 87124, and by the Caritro Deep Learning lab of the ProM facility.

References

- [1] Abhijit Bendale and Terrance Boulton. Towards open world recognition. In *CVPR*, 2015. 2
- [2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 3
- [3] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *ECCV*, 2020. 6
- [4] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *NeurIPS*, 2021. 3
- [5] Pau Panareda Busto, Ahsan Iqbal, and Juergen Gall. Open set domain adaptation for image and action recognition. *TPAMI*, 2018. 1, 2, 3
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 3
- [7] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*, 2019. 1, 3, 6, 7
- [8] Zhuoxiao Chen, Yadan Luo, and Mahsa Baktashmotlagh. Conditional extreme value theory for open set video domain adaptation. In *ACM Multimedia Asia*. 2021. 1, 2, 3, 6, 7
- [9] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *ECCV*, 2020. 1, 3
- [10] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, 2017. 3
- [11] Gabriela Csurka, Timothy M Hospedales, Mathieu Salzmann, and Tatiana Tommasi. Visual domain adaptation in the deep learning era. *Synthesis Lectures on Computer Vision*, 2022. 3
- [12] Victor G Turrissi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *WACV*, 2022. 3
- [13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 6
- [14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 3
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [16] Sepideh Esmailpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI*, 2022. 3, 5, 6, 7
- [17] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 1
- [18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 3
- [19] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 6, 7
- [20] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, 2018. 1
- [21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *JMLR*, 2015. 6, 7
- [22] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019. 3
- [23] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, 2017. 3
- [24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *ICCVW*, 2017. 3
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 7
- [26] Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, 2014. 7
- [27] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 2012. 3
- [28] Amlan Kar, Nishant Rai, Karan Sikka, and Gaurav Sharma. Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In *CVPR*, 2017. 3
- [29] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 3
- [30] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, 2021. 3, 6
- [31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2, 3, 4, 5, 6
- [32] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *IJCV*, 2022. 1, 3

- [33] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 6
- [34] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 1
- [35] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 7
- [36] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 1982. 2
- [37] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 1982. 5
- [38] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2016. 6, 7
- [39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [40] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *ICCVW*, 2019. 3
- [41] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *ACM Multimedia*, 2020. 3
- [42] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 1, 3, 6
- [43] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, 2017. 1
- [44] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 2021. 1
- [45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 6, 7
- [46] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *ICCV*, 2021. 6
- [47] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 7
- [48] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018. 6, 7
- [49] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 2014. 1, 3
- [50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6
- [51] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1
- [52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3
- [54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 3
- [55] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3, 4, 6, 7
- [56] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019. 1
- [57] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 3
- [58] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020. 1