Audio-Adaptive Activity Recognition Across Video Domains

Yunhua Zhang[1] Hazel Doughty[1] Ling Shao[2*] Cees G. M. Snoek[1]

[1]University of Amsterdam [2]Inception Institute of Artificial Intelligence

Abstract

This paper strives for activity recognition under domain shift, for example caused by change of scenery or camera viewpoint. The leading approaches reduce the shift in activity appearance by adversarial training and self-supervised learning. Different from these vision-focused works we leverage activity sounds for domain adaptation as they have less variance across domains and can reliably indicate which activities are not happening. We propose an audio-adaptive encoder and associated learning methods that discriminatively adjust the visual feature representation as well as addressing shifts in the semantic distribution. To further eliminate domain-specific features and include domain-invariant activity sounds for recognition, an audio-infused recognizer is proposed, which effectively models the cross-modal interaction across domains. We also introduce the new task of actor shift, with a corresponding audio-visual dataset, to challenge our method with situations where the activity appearance changes dramatically. Experiments on this dataset, EPIC-Kitchens and CharadesEgo show the effectiveness of our approach. Project page: https://xiaobai1217.github.io/DomainAdaptation.

1. Introduction

The goal of this paper is to recognize activities such as *eating*, *sleeping* or *cutting* under domain shift caused by change of scenery, camera viewpoint or actor, as shown in Figure 1. Existing solutions align distribution-shifted domains inside a single visual video network by adversarial training [5, 20, 27, 29] and self-supervised learning [9, 22, 34]. Although successful, projecting the visual features from different source and target domains into a shared space can make the ability of the model to distinguish between classes in the target domain suffer. We observe that activity sounds can act as natural domain-invariant cues, as they carry rich activity information while exhibiting less variance across domains. We thus propose a video model which adapts to video distribution shifts with the aid of sound.

Many have considered sound in addition to visual analysis for activity recognition within a single domain [18, 24, 25, 28, 31, 38, 39, 45, 46, 49]. For instance, both Gao *et al*. [18] and Korbar *et al*. [24] reduce the computational cost by previewing the audio track, while Lee *et al*. [25] show that combining visual features with audio can better localize actions. However, the cross-modal correspondences become harder to discover when shifting domains, causing existing cross-modal fusion schemes to degrade in performance. Yang *et al*. [48] and Planamente *et al*. [30] propose to directly fuse visual and audio features or predictions for cross-domain activity classification. However, the effectiveness of these methods is reduced when not all activities make a characteristic sound. Different from previous works, we introduce audio-adaptive learning methods and a cross-modal interaction that utilizes the reliable domain-invariant cues within sound to help the video model adapt to the distribution shift.

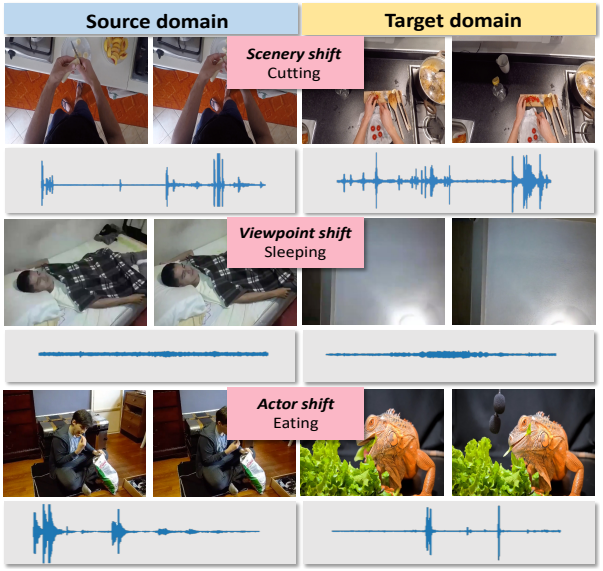

Figure 1. We recognize activities under domain shifts, caused by change of scenery, camera viewpoint or actor, with the aid of sound.

We make three contributions in this paper. First, we propose an audio-adaptive encoder which exploits the rich information from sound to adjust the visual feature representation causing the model to learn more discriminative features in the target domain. This is done by preventing the model from over-fitting to domain-specific visual content, while simultaneously dealing with imbalanced seman-

*Currently at Terminus Group, China.

tic distributions between domains. Second, we introduce an audio-infused recognizer, which eliminates domain-specific features further and allows effective cross-modal interaction across domains by considering domain-invariant activity information within sound. As a third contribution, we introduce the new task of *actor shift*, and a corresponding audio-visual video dataset *ActorShift*, to challenge our approach when the change in actors results in large variation in activity appearance. Experiments on EPIC-Kitchens [12], CharadesEgo [33] and *ActorShift*, demonstrate the advantage of our approach under various video distribution shifts for both audible and silent activities.

2. Related Work

Sound for activity recognition. Many works have utilized sound for within-domain activity recognition in videos, *e.g.*, [18,21,24,25,38,39]. Since there is a natural correlation between the visual and auditive elements of a video, Korbar *et al*. [23] and Asano *et al*. [1] learn audio-visual models in a self-supervised manner. As processing audio signals is much faster than video frames, both Gao *et al*. [18] and Korbar *et al*. [24] reduce computation by previewing the audio track for video analysis. Cross-modal attention is widely used in activity localization [25, 39, 46] and audiovisual video parsing [38, 45] to guide the visual model to focus on the audible regions. Zhang *et al*. [49] conduct repetitive activity counting by using audio signals to decide the sampling rate and predict the reliability of the visual features. As opposed to most works which rely on sound for within-domain activity recognition, we consider its domain-invariant nature for activity recognition across different domains.

Video domain adaptation by vision. The field of vision-focused domain adaptation is extensive (see recent surveys [43, 51]). Here, we focus on video domain adaptation for activity recognition. State-of-the-art visual-only solutions learn to reduce the shift in activity appearance by adversarial training [5, 6, 8, 9, 20, 27, 29] and self-supervised learning techniques [9, 22, 27, 34]. While Jamal *et al*. [20] and Munro and Damen [27] directly penalize domain specific features with an adversarial loss at every time stamp, Chen *et al*. [5], Choi *et al*. [9] and Pan *et al*. [29] attend to temporal segments that contain important cues. Self-supervised learning objectives are also incorporated in [27] and [9] to better align the features across domains by utilizing the correspondences between RGB and optical flow or the temporal order of video clips. Song *et al*. [34] and Kim *et al*. [22] obtain remarkable performance by contrastive learning for self-supervised learning to align the feature distributions between video domains. Instead of relying on the vision modality only, which may present large activity appearance variance, we consider the domain-invariant information within sound to help the model adapt to the visual distribution shift.

Video domain adaptation by vision and audio. As audio signals contain valuable domain-invariant cues, some recent works recognize activities across domains with the aid of sound. Yang *et al*. [48] directly fuse the features from visual and audio modalities before classification. However, this can lead to the visual features dominating the classification since many activities are silent and the audio features are less discriminative. As a result, the complementary information from sound may not be considered. Planamente *et al*. [30] instead align the two modalities with an audio-visual loss. Nonetheless, the audio predictions for silent activities remain unreliable and limit their performance improvements. Instead, we propose audio-adaptive learning that exploits the supervisory signals from sound to adjust to the distribution shift and handle both audible and silent activities.

Additionally, existing datasets *e.g.*, [3, 12, 33, 35] focus on human actors, meaning activities are inherently close in appearance and share commonalities with hand-object interactions. Inspired by the A2D dataset by Xu *et al*. [47], which contains multiple actor classes for activity recognition, we introduce the challenging domain adaptation setting of *actor shift*, in which the shift between humans and animals performing the action results in large appearance and motion differences across domains, further facilitating video domain adaptation by the use of vision and audio.

3. Approach

For activity recognition under domain shift, we consider unsupervised domain adaptation where we have: a set of labeled source videos $\mathcal{S}=\{(X_1^{\mathcal{S}}, y_1^{\mathcal{S}}), \ldots, (X_N^{\mathcal{S}}, y_N^{\mathcal{S}})\}$ and a set of unlabeled target videos $\mathcal{T}=\{X_1^{\mathcal{T}}, \ldots, X_M^{\mathcal{T}}\}$. In each domain, X and y indicate a video sample and the corresponding activity class label, while N and M are the number of samples in the source and target domain. Using all available training data from the source and the target domains, the task is to train an activity recognition model, which performs well on (unseen) videos from the target domain.

We train our audio-adaptive model in two stages using videos from source and target domains with accompanying audio. In the first stage we train our audio-adaptive encoder (Section 3.1) that uses audio to adapt a visual encoder to be more robust to distribution shifts. In the second stage we train our audio-infused recognizer (Section 3.2) using pseudo-labels from the audio-adaptive encoder for the target domain and the ground-truth labels for the source domain. The audio-infused recognizer maps the source and target domains into a common space and fuses audio and visual features to produce an activity prediction for either domain.

3.1. Stage 1: Audio-Adaptive Encoder

Our audio-adaptive encoder $\mathcal{E}(\cdot)$, detailed in Figure 2, consists of a visual encoder $\mathcal{V}(\cdot)$, an audio encoder $\mathcal{A}(\cdot)$ and an audio-based attention module $\psi(\cdot)$. Since the sounds of activities have less variance across domains, $\mathcal{E}(\cdot)$ aims

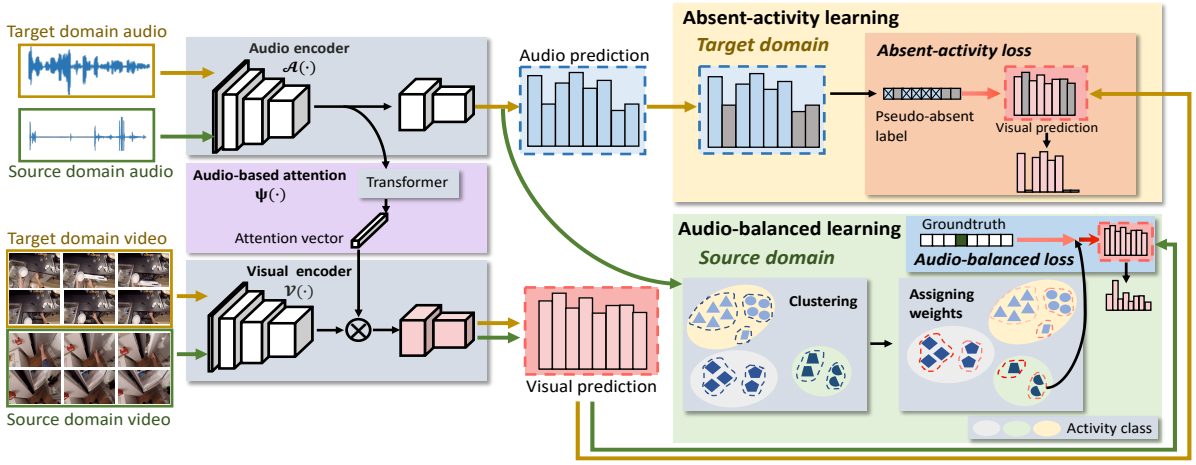

Figure 2. **Audio-adaptive encoder for activity recognition under domain shift.** With a pre-trained audio encoder, we train the visual encoder and audio-based attention module, which guides the visual encoder to focus on the activity relevant features. We do this with two audio-adaptive learning methods: absent-activity learning and audio-balanced learning. The absent activity learning operates in the target domain and uses the audio predictions to indicate which activities cannot be heard in the video. The visual predictions are then encouraged to have low probabilities for these 'pseudo-absent' activities. The audio-balanced learning uses audio in the source domain to cluster samples in each activity class into clusters according to the sounds of the object/environment interacted with. In the audio-balanced loss the rare activities and interactions are weighted higher to handle the semantic shift between domains.

to extract visual features that are invariant but discriminative under domain shift with the aid of $\mathcal{A}(\cdot)$ pre-trained for audio-based activity recognition. To this end, we train $\mathcal{V}(\cdot)$ and $\psi(\cdot)$ with two audio-adaptive learning methods: absent-activity learning for unlabeled target data and audio-balanced learning for labeled source data. The former aims to remove irrelevant parts of the visual features while the latter helps to handle the differing label distribution between domains. Once trained, for each video, we can extract an audio feature vector from $\mathcal{A}(\cdot)$ and a series of visual features from $\mathcal{V}(\cdot)$ with which to train our audio-infused recognizer (Section 3.2) for activity classification.

Audio-based attention. We use an audio-based attention module $\psi(\cdot)$ to adapt the visual encoder to focus on activity-relevant features. For example, the visual model may predict the activity *washing* because of the presence of a sink. However, without the sound of water the attention module suppresses the channels encoding the sink thus increasing the prediction of the correct class. The attention module is based on the transformer encoder [13, 14, 42]. It takes the audio features as input and outputs the channel attention feature vector, which is multiplied with the visual features.

Absent-activity learning. The absent-activity learning uses audio in the target domain to train the attention module and visual encoder. Naively, we could treat the class with the highest probability from the visual encoder as the pseudo label. However, doing so can create biased pseudo-labels as irrelevant objects often appear in a scene. Instead, we use the audio predictions to guide the visual pseudo-labels. While we may not be confident which activity is happening in a video, particularly for silent videos, we can often be confident that certain activities with distinctive sounds are *not* occurring in a video. We call these "absent activities". To learn from these absent activities, we generate pseudo-absent labels for the unlabeled target domain videos, which indicate the activities with the lowest probabilities from the audio encoder. The visual encoder is then encouraged to predict these unlikely classes with low probability.

Specifically, for an unlabeled video $X^{\mathcal{T}}$ in the target domain, we obtain the audio-based activity probability distribution $\mathbf{p}_a^{\mathcal{T}} \in \mathbb{R}^K$ (K is the number of classes) from the audio encoder $\mathcal{A}(\cdot)$ trained on labeled source data. From this we obtain the set of absent activities $\mathcal{Q}$ by taking the lowest r predictions in $\mathbf{p}_a^{\mathcal{T}}$, *i.e.*, the classes with the lowest probabilities from the audio encoder. We also extend this to multi-label classification by instead assuming the $(1-\alpha_k)\gamma$ percent videos with the lowest probabilities do not contain class k, where $\gamma \in (0, 1]$ and α_k is the percentage of videos containing each activity class in the labeled source domain.

Our loss for absent-activity learning is formulated as:

$$l_{\mathrm{A}}(\mathbf{p}_v^{\mathcal{T}}, \mathcal{Q}) = -\sum_{q \in \mathcal{Q}} \log(1 - p_{v,q}^{\mathcal{T}}), \tag{1}$$

where $p_{v,q}^{\mathcal{T}}$ is the probability output for the qth class for the video $X^{\mathcal{T}}$. With this loss, the visual encoder is able to ignore confounding visual features and generate less-noisy pseudo-labels for the target domain. This allows our model to better capture high-level semantic information between domains based on both appearance and motion cues.

Audio-balanced learning. Besides a change in visual appearance, domain shift can also be caused by a

change in label distributions [27] and frequencies of objects/environments. For example, the *open* activity may commonly occur on a 'cupboard' in the source domain but be more common with a 'can' in the target. These two cases result in different audio-visual activity appearances. We address such challenges with our audio-balanced learning, which not only handles imbalance in activity classes, but also imbalance in terms of the objects or the environment being interacted with.

To this end, we first use k-means to group the video samples inside each activity class by their audio feature $\mathbf{f}_a^{\mathcal{S}}$ with the assumption that each group represents a different type of object or environment. We use audio features for clustering as they can indicate the material of the interacted objects or the environment the action is performed in, while being invariant to appearance changes. The number of interaction clusters per activity class is determined by the Elbow method [37], which favours a small number while obtaining a low ratio of dispersion both between and within clusters.

We based our *audio-balanced* loss on the class-balanced loss by Cui *et al*. [11]. When using the original class-balanced loss on a source domain video $X^{\mathcal{S}}$ with visual probabilities $\mathbf{p}_v^{\mathcal{S}}$ we can balance over our activity classes:

$$l_{\text{CB}}(\mathbf{p}_v^{\mathcal{S}}, y^{\mathcal{S}}) = \frac{1-\beta}{1-\beta^{n_y}} \mathcal{L}(\mathbf{p}_v^{\mathcal{S}}, y^{\mathcal{S}}), \tag{2}$$

where $\mathcal{L}$ is a classification loss, *e.g*., softmax cross-entropy loss and n_y is the number of training samples of ground-truth activity class y. $\beta \in [0, 1)$ is a hyper-parameter which controls the weighting factor $\frac{1-\beta}{1-\beta^{n_y}}$. As $\beta \to 1$, this weighting factor becomes inversely proportional to the effective number of samples inside each class so that tail classes in the source domain are weighted higher in training.

With our *audio-balanced loss* we include an additional weighting factor so the long tail of object interactions are also accounted for with our interaction clusters:

$$l_{\text{B}}(\mathbf{p}_v^{\mathcal{S}}, y^{\mathcal{S}}) = \frac{1-\beta}{1-\beta^{n_{y,j}}} l_{\text{CB}}(\mathbf{p}_v^{\mathcal{S}}, y^{\mathcal{S}}). \tag{3}$$

$n_{y,j}$ is the number of samples for the jth interaction cluster that video $X^{\mathcal{S}}$ is assigned within ground-truth activity $y^{\mathcal{S}}$. By this loss, both rare activities and rare interactions from frequent activities are given a high weight during training. This means the classifier can generalize well to the target domain where the distribution of activities and interactions may not be the same.

Audio-adaptive encoder loss. The absent-activity loss and the audio-balanced loss are combined to obtain the overall loss for training the visual encoder $\mathcal{V}(\cdot)$ and audio-based attention $\psi(\cdot)$ inside the audio-adaptive encoder $\mathcal{E}(\cdot)$:

$$l_{\mathcal{E}} = \sum_{(X_i)\in\mathcal{T}} l_{\text{A}}(\mathbf{p}_{i,v}^{\mathcal{T}}, \mathcal{Q}_i) + \sum_{(X_j,y_j)\in\mathcal{T}} l_{\text{B}}(\mathbf{p}_{j,v}^{\mathcal{S}}, y_j^{\mathcal{S}}). \tag{4}$$

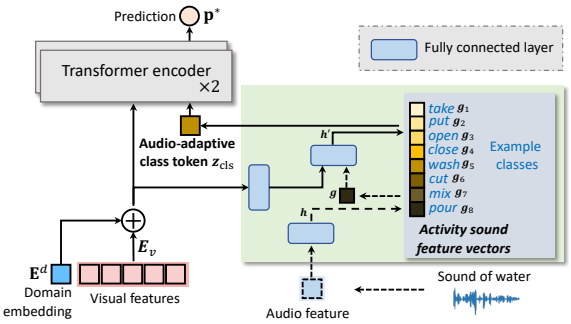

Figure 3. **Audio-infused recognizer.** We add domain embedding E_d to encourage a common visual representation across domains. Then, an audio-adaptive class token is obtained from a series of activity sound feature vectors, considering both audio and visual features. It is sent into the transformer together with the visual features. By the transformer's self attention, this token aggregates information from visual features with the domain-invariant audio activity cues for activity classification.

3.2. Stage 2: Audio-Infused Recognizer

While audio can help focus on the activity-relevant visual features, there is still a large difference between the appearance of activities in different domains. To further eliminate domain-specific visual features and fuse the activity cues from the audio and visual modalities we propose the audio-infused recognizer $\mathcal{R}(\cdot)$, visualized in Figure 3.

Transformer with domain embedding. We adopt a transformer encoder since its core mechanism, self-attention, can efficiently encode multi-modal representations [16, 36, 50]. For a vanilla version, we take the input sequence:

$$\mathbf{z}^m = [\mathbf{z}_{\text{cls}}^m; \mathbf{f}_{v,1}\mathbf{E}_v; \cdots; \mathbf{f}_{v,n}\mathbf{E}_v; \mathbf{f}_{a,1}\mathbf{E}_a, ; \cdots; \mathbf{f}_{a,n}\mathbf{E}_a], \tag{5}$$

where $\mathbf{z}_{\text{cls}}^m$ is the learnable class token defined as in [14], and $\{\mathbf{f}_{v,1}, \cdots, \mathbf{f}_{v,n} | \mathbf{f}_{v,\cdot} \in \mathbb{R}^{C_v}\}$ and $\{\mathbf{f}_{a,1}, \cdots, \mathbf{f}_{a,n} | \mathbf{f}_{a,\cdot} \in \mathbb{R}^{C_a}\}$ are the visual and audio features of n clips from video X. $\mathbf{E}_v \in \mathbb{R}^{C_v \times D}$ and $\mathbf{E}_a \in \mathbb{R}^{C_a \times D}$ are linear projections to map the visual and audio features to D dimensions. To map source and target domains into a common space, we first learn a domain embedding $\mathbf{E}^d \in \mathbb{R}^D$ ($d \in \{\mathcal{S}, \mathcal{T}\}$), which contains both positive and negative values and is added to suppress domain-specific visual features. Then, the input sequence for the transformer becomes:

$$\mathbf{z}' = [\mathbf{z}_{\text{cls}}^m; \mathbf{f}_{v,1}\mathbf{E}_v + \mathbf{E}^d; \cdots; \mathbf{f}_{v,n}\mathbf{E}_v + \mathbf{E}^d; \mathbf{f}_{a,1}\mathbf{E}_a, ; \cdots; \mathbf{f}_{a,n}\mathbf{E}_a]. \tag{6}$$

Audio-adaptive class token. Ideally, the transformer's self attention will aggregate audio and visual features with the class token to predict the correct activity. However, the cross-modal correspondences are difficult to find under distribution shift, meaning the prediction may rely on the more discriminative, but less domain-invariant, visual features. To address this, we propose to generate an audio-adaptive class token, which is initialized from the audio activity class prediction and gradually aggregates the visual features while

		Source Domain Setting		Target Domain Setting		
Shift	**Video Dataset**	*Source Domain*	*Train*	*Target Domain*	*Train*	*Test*
Scenery	EPIC-Kitchens-55 [12]	Kitchens	7,935	Kitchens	7,935	2,114
Viewpoint	CharadesEgo [33]	Third-person view	3,083	Ego-centric view	3,083	825
Actor	ActorShift (*ours*)	Human actors	1,305	Animal actors	35	165

Table 1. **Domain adaptation benchmarks for activity recognition under scenery, viewpoint and actor shift** with the datasets used and number of videos per source and target split. Scenery and viewpoint shift are present in existing datasets. We propose the actor shift setting and dataset to tackle the challenge of a severe change in activity appearance. The dataset is available on the project website.

keeping its own audio-based activity information through the transformer. As shown in Figure 3, the audio-adaptive class token is obtained from a series of activity sound vectors $\{\mathbf{g}_k \in \mathbb{R}^D\}_{k=1}^K$, with each representing an activity class. They capture global context information and serve as the representation bottleneck to provide regularization for model learning [2, 32]. For selection, the feature vector from the audio adaptive encoder $\mathcal{A}(X)$ is first processed by a fully connected layer to give the activity probabilities $\mathbf{h} \in \mathbb{R}^K$. Then, an initial vector is obtained by $\mathbf{g}{=}\sum_{k=1}^K h_k * \mathbf{g}_k$. We include visual features to help silent activities select the representative vector. To avoid the visual features dominating, we project them to a lower dimension with a fully connected layer before concatenating them with the initial vector $\mathbf{g}$. The concatenated vector is given to another fully connected layer which outputs the probabilities $\mathbf{h}'$ for each type of activity sound. Finally, we obtain the audio representation $\mathbf{z}_{\text{cls}}{=}\sum_{k=1}^K h'_k * \mathbf{g}_k$, which serves as the class token. Consequently, the input sequence for the transformer becomes:

$$\mathbf{z}{=}[\mathbf{z}_{\text{cls}}; \mathbf{f}_{v,1}\mathbf{E}_v + \mathbf{E}^d, ; \cdots ; \mathbf{f}_{v,n}\mathbf{E}_v + \mathbf{E}^d], \qquad (7)$$

where $\mathbf{z}_{\text{cls}}$ is the audio-adaptive class token. The class token output state is further sent to a fully connected layer to get the final prediction $\mathbf{p}^*$. For audible activities, the activity sound vector can be accurately selected and kept discriminative for audiovisual interaction. For silent activities, the vector is obtained from environmental sound, which indicates the presence of multiple possible activities. The vector becomes more discriminative as the transformer progressively enhances it through the visual features.

Audio-infused recognizer loss. We train the audio-infused recognizer on both source and target videos with the loss:

$$l_{\mathcal{R}} = \sum_{(X_i,y_i)\in\{\mathcal{S},\mathcal{T}\}} \mathcal{L}(\mathbf{p}_i^*, y_i) + \eta\Big(\mathcal{L}(\mathbf{h}_i, y_i) + \mathcal{L}(\mathbf{h}'_i, y_i)\Big), \qquad (8)$$

where hyperparameter η balances the loss terms and y_i is the groundtruth or, in the case of the unlabeled video, the hard pseudo-label. $\mathbf{p}_i^*$ is the final classification prediction, and $\mathbf{h}_i$ and $\mathbf{h}'_i$ are the probabilities for the activity sound vectors outputted by the first and second fully connected layers. The first term $\mathcal{L}(\mathbf{p}_i^*, y_i)$ optimizes the transformer to predict the correct activity class, while the second term $\mathcal{L}(\mathbf{h}_i, y_i) + \mathcal{L}(\mathbf{h}'_i, y_i)$ optimizes the activity sound vectors. We are now ready to validate the effectiveness of our approach on three domain adaptation benchmarks as highlighted in Figure 1, summarized in Table 1 and detailed next.

4. Domain Adaptation Benchmarks

Scenery shift. We study scenery shift in the ***EPIC-Kitchens-55*** [12] dataset, which contains first-person videos of fine-grained kitchen activities. The domain adaptation benchmark proposed by Munro and Damen [27] uses three domain partitions (D1, D2 and D3), where each domain is a different person in a different kitchen. The task is to adapt between each pair of domains. This benchmark focuses on eight activity classes (verbs), which occur in combination with different objects, with a severe class imbalance. The kitchens have different appearances and contain different utensils.

Viewpoint shift. We consider viewpoint shift in the ***CharadesEgo*** dataset by Sigurdsson *et al*. [33]. It contains paired videos of the same activities, recorded from first and third-person perspective. It has 3,083 and 825 videos per viewpoint for training and testing, spanning 157 activity classes. Following [8], we treat the third-person videos as the source domain and the first-person videos as the target domain. The changing views make the activities appear visually different, resulting in a large domain gap.

Actor shift. While both EPIC-Kitchens and CharadesEgo contain considerable domain shifts, there are still some inherent similarities between the domains in these datasets. Since all the actors are humans, latent signals describing the way hands and objects interact are shared between domains. Therefore, we introduce an even more challenging domain shift setting to further facilitate video domain adaption research and demonstrate the potential of our method. We introduce *ActorShift*, where the domain shift comes from the change in actor species: we use humans in the source domain and animals in the target domain. This causes large variances in the appearance and motion of activities.

For the corresponding dataset we select 1,305 videos of 7 human activity classes from Kinetics-700 [3] as the source domain: *sleeping*, *watching tv*, *eating*, *drinking*, *swimming*, *running* and *opening a door*. For the target domain we collect 200 videos from YouTube of animals performing the same activities. We divide them into 35 videos for training (5 per class) and 165 for evaluation. The target domain data is

scarce, meaning there is the additional challenge of adapting to the target domain with few unlabeled examples.

Evaluation criteria. Following standard practice [27, 33], we report top-1 accuracy on EPIC-Kitchens and ActorShift for single-label classification, and mAP (mean average precision) on CharadesEgo for multi-label classification.

5. Results

We first describe the implementation details before ablating the components of our method and comparing to prior works for each type of domain shift.

Implementation details. For our visual encoder $\mathcal{V}(\cdot)$ we use SlowFast [15], unless stated otherwise. For the audio encoder $\mathcal{A}(\cdot)$ we use ResNet-18 [19]. The audio-based attention module $\psi(\cdot)$ consists of eight transformer encoder layers [14] with a final fully connected layer to obtain the attention vector for the visual encoder. The inputs are intermediate audio features from $\mathcal{A}(\cdot)$ (conv3) along with a learnable class token defined as in [14] (note this is different from our audio-adaptive class token used in $\mathcal{R}(\cdot)$). The output state of the class token passes through the fully connected layer to obtain the attention vector for $\mathcal{V}(\cdot)$. We set the parameters of our absent activity loss to r=3, γ=0.05 and β=0.999. Our audio-infused recognizer $\mathcal{R}(\cdot)$ consists of two transformer encoder layers [14] and three fully connected layers for generating the class token. The sequence dimension D is 512 and each layer has 8 self-attention heads. More details are in the supplementary.

5.1. Ablation Study

For ablations we use RGB and audio modalities on both EPIC-Kitchens and CharadesEgo. During training, all labeled source videos are used. With EPIC-Kitchens all target videos are unlabelled, while for CharadesEgo we use half labelled and half unlabelled for semi-supervised domain adaptation as in [8]. Since EPIC-Kitchens contains multiple adaptation settings, we report the average. Ablations on component internals are provided in the supplementary.

Stage 1: Audio-adaptive encoder. We report results in Table 2. We first consider the audio-adaptive encoder alone. Initially, we train only the visual encoder with a standard softmax cross-entropy loss on the source domain. Simply generating channel attention for the visual features with our audio-based attention module already improves performance by 3.2% top-1 accuracy on EPIC-Kitchens and 0.4% mAP on CharadesEgo. Since audio contains useful activity information, this attention helps the visual encoder focus on relevant features. Adding the absent-activity learning results in 2.5% and 0.9% improvements, demonstrating that the pseudo-absent labels increase the discriminative ability of the model in the target domain. We observe that adopting the audio-balanced learning and replacing the softmax cross-entropy with our audio-balanced loss delivers a further 2.0% and 0.6% increase. This highlights the importance of addressing the label distribution shift in domain adaption.

Model	EPIC-Kitchens Top-1 (%) ↑	CharadesEgo mAP (%) ↑
Stage 1: Audio-adaptive encoder $\mathcal{E}(\cdot)$		
Visual encoder $\mathcal{V}(\cdot)$	48.0	23.1
+ Audio-based attention $\psi(\cdot)$	51.2	23.5
+ Absent-activity learning	53.7	24.4
+ Audio-balanced learning	55.7	25.0
Stage 2: Audio-infused recognizer $\mathcal{R}(\cdot)$		
+ Vanilla multi-modal transformer $\mathbf{z}^m$	56.1	25.0
+ Domain embedding $\mathbf{z}'$	57.2	25.4
+ Audio-adaptive class token $\mathbf{z}$	59.2	26.3

Table 2. **Model components ablation.** All components in the audio-adaptive encoder and the audio-infused recognizer contribute to performance improvement under distribution shift. For both EPIC-Kitchens and CharadesEgo the improvements over a vanilla SlowFast visual encoder are considerable.

Model	Activities Silent	Activities Audible	Overall mAP (%) ↑
Visual encoder $\mathcal{V}(\cdot)$	23.2	22.7	23.1
Full model	26.3	25.9	26.3

Table 3. **Benefit over silent and audible activities** on CharadesEgo. Our audio-adaptive model benefits both activity types.

Stage 2: Audio-infused recognizer. For the audio-infused recognizer, we first consider a vanilla transformer. It takes as input $\mathbf{z}^m$ (Eq. 5), *i.e*. the audio and visual features from the audio-adaptive encoder, mapped by $\mathbf{E}_v$ and $\mathbf{E}_\mathbf{a}$ into a common space, alongside a learnable class token. This only gives a marginal improvement in results. Adding the domain embedding $\mathbf{E}^d$ to reduce domain-specific visual features in $\mathbf{z}'$ (Eq. 6) gives a benefit of 1.1% on EPIC-Kitchens and 0.4% on CharadesEgo. This is because the cross-modal correspondences become easier to discover. When we replace the plain audio features and single learnable class token with our audio-adaptive class token to get $\mathbf{z}$ (Eq. 7), we observe further improvements of 2.0% and 0.9%. This is expected, as the audio-adaptive class token better incorporates complementary information from sound for the final activity classification, with a standard learnable class token the visual features will dominate the fusion inside the transformer.

Benefit for silent activities. In Table 3, we demonstrate the effect of our full model on silent and audible activities separately. We focus on CharadesEgo since only 13 out of 157 classes have a characteristic sound (see supplementary). Our model obtains ∼3% absolute increase for both silent and audible activities over a visual-only encoder. We conclude that audio is helpful for handling visual distribution shifts even for activities which do not have a characteristic sound.

Benefit for silent videos. We have also tested our approach when the audio track is available for training but unavailable during inference. On EPIC-Kitchens, the audio-adaptive encoder achieves 50.7% top-1 accuracy, still an improvement

Method	Modality			EPIC-Kitchen Activity Recognition Across Domains						
	RGB	Flow	Audio	D2 → D1	D3 → D1	D1 → D2	D3 → D2	D1 → D3	D2 → D3	*Mean*
I3D backbone										
Source-only [27]	✓	✓		42.5	44.3	42.0	56.3	41.2	46.5	45.5
Munro and Damen [27]	✓	✓		48.2	50.9	49.5	56.1	44.1	52.7	50.3
Planamente *et al*. [30]†	✓	✓	✓	48.5	50.9	49.7	56.3	44.8	52.5	50.5
Yang *et al*. [48]†	✓	✓	✓	49.2	51.0	49.8	56.5	45.7	52.3	50.8
Kim *et al*. [22]	✓	✓		49.5	51.5	50.3	56.3	46.3	52.0	51.0
Song *et al*. [34]	✓	✓		49.0	**52.6**	52.0	55.6	45.5	52.5	51.2
This paper	✓	✓	✓	**51.9**	48.7	**53.2**	**63.2**	**52.1**	**55.5**	**54.1**
SlowFast backbone										
This paper	✓	✓	✓	**59.3**	**59.1**	**59.5**	**69.1**	**54.8**	**64.3**	**61.0**

† Based on our re-implementation using our features for RGB, flow and audio.

Table 4. **Activity recognition under scenery shift** on EPIC-Kitchens for the unsupervised domain adaptation setting. Our audio-adaptive model achieves state-of-the-art top-1 accuracy, and benefits from audio more than the audio-visual fusion methods used in prior works [30,48]. Results increase further with a SlowFast backbone. More comparisons and modality-combinations are provided in the supplementary.

over visual encoder only (48.0%). With both the audio-adaptive encoder and audio-infused recognizer, the result improves to 51.2%. This indicates our approach effectively uses audio to help the visual encoder learn a more discriminative feature representation in the target domain, even when audio is absent during inference.

Benefit for the long-tail. In Figure 4, we demonstrate the benefit of audio-balanced learning towards activities that are rare in the source domain but are more frequent in the target domain. We use EPIC-Kitchens since it contains a long-tail of different object interactions (nouns) in each activity class (verb). We treat verb-noun pairs as frequent when they occur more than 10 times in the source domain, else they are considered rare. As the distribution of activities (verbs) changes across domains, the class-balanced loss [11] improves over the standard softmax cross-entropy loss. However, the domain shift also causes imbalance in the distribution of interactions (nouns). Because we balance the loss of each pseudo-interaction by clustering, our audio-balanced loss is especially helpful for the rare interactions (0-1 and 2-10 instances) where it obtains ∼3.5% improvement. In comparison to the class-balanced loss we are slightly worse on frequent interactions, as we give higher weight to less common interactions. As interactions have a long-tail, our audio-balanced loss does result in an overall improvement.

5.2. Comparison with State-of-the-Art

Scenery shift. We first demonstrate the effectiveness of our approach for domain adaptation on EPIC-Kitchens, as defined by Munro and Damen [27]. Here, different domains mean a change in scenery. The results are shown in Table 4. We first note that our approach gives ∼3% improvement over the best performing prior works with the same I3D backbone. A further ∼7% improvement can be gained from using SlowFast as the backbone [15]. There are several reasons for this improvement. First, our model utilizes the domain-invariant nature of audio signals to produce reliable pseudo-absent labels for the target domain video during training. This is particularly helpful for first-person videos where the activity may happen out of view. In addition, both RGB and Flow suffer from large appearance variance making it harder to guide domain-adaption through these modalities alone. Second, since the dataset has imbalanced label distributions, treating all the classes and interactions equally, as in prior works, results in inaccurate predictions when the semantic distribution shifts.

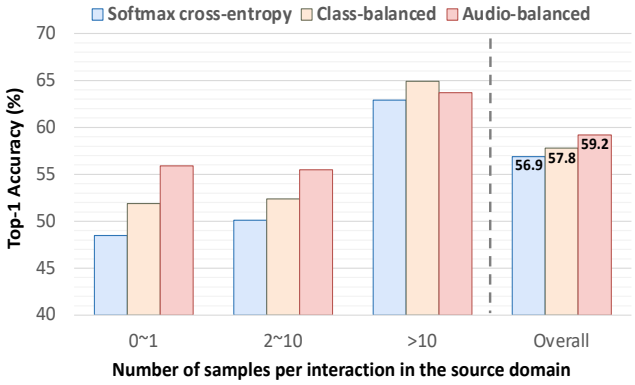

Figure 4. **Benefit for the long-tail** on EPIC-Kitchens. Our audio-balanced loss learns rare activities in the source domain to generalize better to unknown activity distributions in the target domain.

We also compare our full model with alternative audio-visual approaches proposed for cross-domain activity recognition [30,48]. We let both of them use the same inputs, *i.e*., the features as outputted by the visual and audio encoders. Both of them use an adversarial loss to first align the visual features between domains and fuse visual and audio features or predictions afterwards. This causes the visual features to dominate the classification while the complementary information from sound may not be considered. Planamente *et al*. [30] introduce an audio-visual loss, so the two modalities make a more balanced contribution towards the prediction. However, the audio predictions for silent activities are unreliable and harm their accuracy. Our model better combines the complementary information in the audio and visual modali-

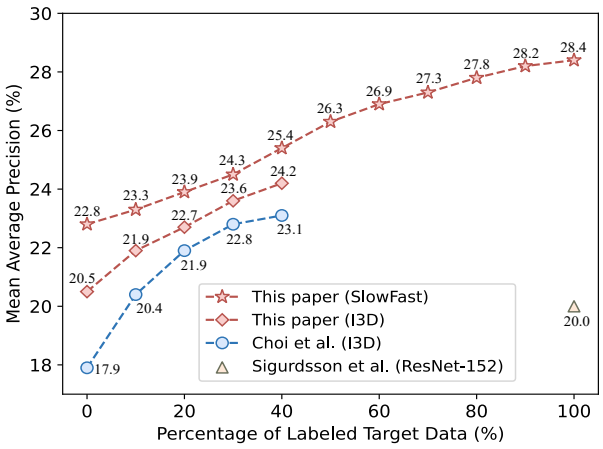

Figure 5. **Activity recognition under viewpoint shift** on CharadesEgo. Using all source data, we compare with [8, 33] under varying amounts of labeled target training data. Our model obtains favourable results under all settings.

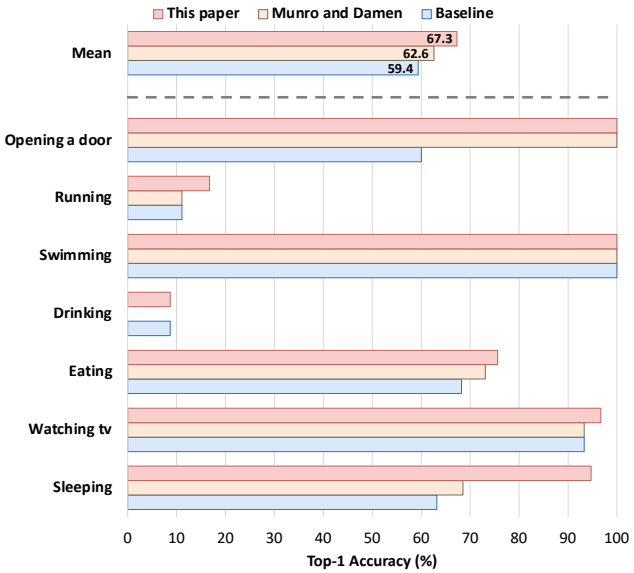

Figure 6. **Activity recognition under actor shift** on our ActorShift dataset. When the visual similarities for the same activity are difficult to discover between domains, our model can use additional cues from sound to improve the recognition accuracy.

ties, effectively coping with many activities being silent.

Viewpoint shift. In this comparison we consider viewpoint shift in CharadesEgo [33], following the semi-supervised setting of Choi *et al*. [8]. Meaning we have some labeled target domain videos available during training. The results are shown in Figure 5. Our method achieves better results than Choi *et al*. [8] with the same I3D RGB backbone [3], for all amounts of labeled target videos. When adopting the SlowFast RGB backbone [15], we again further improve performance for all settings. In the supplementary, we also provide a favorable comparison with Li *et al*. [26] under their fully-supervised setting. Since CharadesEgo contains paired first-person and third-person videos, we can test whether our method needs to see the same action instance from different viewpoints as in previous methods [33] or whether it can make use of unpaired videos. When half of the paired videos from both views are used, we achieve a mAP of 29.9. When we use unpaired videos, the performance remains unchanged. We conclude our approach does not require paired training videos to be robust to viewpoint shift.

Actor shift. For this experiment, we use our ActorShift dataset and compare our model with the method by Munro and Damen [27], as their code is available. For fair comparison, we replace their I3D backbone with the same SlowFast backbone used for our model. We also show a baseline of the SlowFast model trained on source domain video only. The results are shown in Figure 6. While the method proposed by Munro and Damen [27] achieves good performance, our audio-adaptive approach better handles the large activity appearance variance caused by the shift in actors. For example, humans and animals sleep in visually different places and positions, while the sound of snoring or breathing is common to both. All models struggle with silent activities when there is both a large shift in appearance and a significant difference in sounds of activities between the domains, such as *drinking* and *running*. We provide examples in the supplemental material, which are of interest for future work.

6. Discussion

Limitations. During training, our method needs videos from both source and target domains, and all should have an audio track with decent quality, limiting our approach to multimodal video training sets. While audio at test-time is not required, it benefits activity recognition results considerably.

Potential negative impact. When deployed our approach will have to record, store and process video and audio information related to human activities, which will have privacy implications for some application domains.

Conclusions. We propose to recognize activities under domain shift with the aid of sound, using a novel audiovisual model. By leveraging the domain-invariant activity information within sound, our model improves over both silent and audible activities as well as rare activities in the source domain. Experiments on two domain adaptation benchmarks demonstrate that our approach has better adaptation ability than visual-only solutions and benefits from audio more than alternative audiovisual fusion methods used in prior works. We also show that our model better handles large activity appearance variance caused by the shift in actors.

Acknowledgement. This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

References

[1] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020.

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017.

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: a large-scale audio-visual dataset. In *ICASSP*, 2020.

[5] 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.

[6] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *CVPR*, 2020.

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

[8] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *WACV*, 2020.

[9] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *ECCV*, 2020.

[10] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020.

[11] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

[12] 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.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020.

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[18] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[20] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, 2018.

[21] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019.

[22] 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.

[23] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018.

[24] Bruno Korbar, Du Tran, and Lorenzo Torresani. SCSampler: Sampling salient clips from video for efficient action recognition. In *ICCV*, 2019.

[25] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *ICLR*, 2021.

[26] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021.

[27] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020.

[28] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021.

[29] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *AAAI*, 2020.

[30] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Cross-domain first person audio-visual action recognition through relative norm alignment. *arXiv preprint arXiv:2106.01689*, 2021.

[31] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *CVPR*, 2021.

[32] Alexandre Rame and Matthieu Cord. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *ICLR*, 2021.

[33] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018.

[34] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *CVPR*, 2021.

[35] 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.

[36] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

[37] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[38] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: weakly-supervised audio-visual video parsing. In *ECCV*, 2020.

[39] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

[40] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019.

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[43] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[44] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

[45] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 2021.

[46] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019.

[47] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? Action understanding with multiple classes of actors. In *CVPR*, 2015.

[48] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. EPIC-KITCHENS-100 unsupervised domain adaptation challenge for action recognition 2021: Team M3EM technical report. *arXiv preprint arXiv:2106.10026*, 2021.

[49] Yunhua Zhang, Ling Shao, and Cees GM Snoek. Repetitive activity counting by sight and sound. In *CVPR*, 2021.

[50] Linchao Zhu and Yi Yang. ActBERT: Learning global-local video-text representations. In *CVPR*, 2020.

[51] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.