Audio-Adaptive Activity Recognition Across Video Domains

Zhang, Y.; Doughty, H.; Shao, L.; Snoek, C.G.M.

DOI
10.48550/arXiv.2203.14240
10.1109/CVPR52688.2022.01342

Publication date
2022

Document Version
Author accepted manuscript

Published in
2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Audio-Adaptive Activity Recognition Across Video Domains

Yunhua Zhang\textsuperscript{1}  Hazel Doughty\textsuperscript{1}  Ling Shao\textsuperscript{2*}  Cees G. M. Snoek\textsuperscript{1}

\textsuperscript{1}University of Amsterdam  \textsuperscript{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: \url{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 [18,24,25,31,38,39,45,46,49]. For instance, both Gao \textit{et al.} [18] and Korbar \textit{et al.} [24] reduce the computational cost by previewing the audio track, while Lee \textit{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 \textit{et al.} [48] and Planamente \textit{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.

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-
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 auditory 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 $S = \{(X_1^S, y_1^S), \ldots, (X_N^S, y_N^S)\}$ and a set of unlabeled target videos $T = \{X_1^T, \ldots, X_M^T\}$. 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 $E(\cdot)$, detailed in Figure 2, consists of a visual encoder $V(\cdot)$, an audio encoder $A(\cdot)$ and an audio-based attention module $\psi(\cdot)$. Since the sounds of activities have less variance across domains, $E(\cdot)$ aims
to extract visual features that are invariant but discriminative under domain shift with the aid of $A(\cdot)$ pre-trained for audio-based activity recognition. To this end, we train $V(\cdot)$ and $\psi(\cdot)$ with two audio-adaptive learning methods: absent-activity learning and audio-balanced 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 $A(\cdot)$ and a series of visual features from $V(\cdot)$ 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 $X^T$ in the target domain, we obtain the audio-based activity probability distribution $p^T_a \in \mathbb{R}^K$ ($K$ is the number of classes) from the audio encoder $A(\cdot)$ trained on labeled source data. From this we obtain the set of absent activities $Q$ by taking the lowest $r$ predictions in $p^T_a$, 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 $\alpha_k$ is the percentage of videos containing each activity class in the labeled source domain.

Our loss for absent-activity learning is formulated as:

$$l_A(p^T_a, Q) = - \sum_{q \in Q} \log(1 - p^T_{a,q}),$$

where $p^T_{a,q}$ is the probability output for the $q$th class for the video $X^T$. 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 $f_a^S$ 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 $X^S$ with visual probabilities $p_v^S$ we can balance over our activity classes:

$$l_{CB}(p_v^S, y^S) = \frac{1 - \beta}{1 - \beta^{n_y}} l(C(p_v^S, y^S),$$

(2)

where $l$ 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 $\beta \rightarrow 1$, 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:

$$l_{ICB}(p_v^S, y^S) = \frac{1 - \beta}{1 - \beta^{n_{y,j}}} l_{CB}(p_v^S, y^S).$$

(3)

$n_{y,j}$ is the number of samples for the $j$th interaction cluster that video $X^S$ is assigned within ground-truth activity $y^S$. 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 $V(\cdot)$ and audio-based attention $\psi(\cdot)$ inside the audio-adaptive encoder $E(\cdot)$:

$$l_E = \sum_{(X,v) \in T} l_A(p_v^T, Q_v) + \sum_{(X,y_j) \in T} l_{ICB}(p_v^S, y_j).$$

(4)

**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 $R(\cdot)$, 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:

$$z^{m} = [z_{ch}^{m}; f_{v,1} E_v; \cdots; f_{v,n} E_v; f_{a,1} E_a; \cdots; f_{a,n} E_a],$$

(5)

where $z_{ch}^{m}$ is the learnable class token defined as in [14], and $\{f_{v,1}, \cdots, f_{v,n}\} \in \mathbb{R}^{C_x}$ and $\{f_{a,1}, \cdots, f_{a,n}\} \in \mathbb{R}^{C_a}$ are the visual and audio features of $n$ clips from video $X$. $E_v \in \mathbb{R}^{C_x \times D}$ and $E_a \in \mathbb{R}^{C_a \times D}$ 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 $E_d \in \mathbb{R}^{D}$ ($d \in \{S, T\}$), which contains both positive and negative values and is added to suppress domain-specific visual features. Then, the input sequence for the transformer becomes:

$$z' = [z_{ch}^{m}; f_{v,1} E_v + E_d; \cdots; f_{v,n} E_v + E_d; f_{a,1} E_a; \cdots; f_{a,n} E_a].$$

(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
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 \( \{g_k \in \mathbb{R}^D\}_{k=1}^K \), with each representing an activity class.
They capture global context information and serve as the rep-
resentation bottleneck to provide regularization for model
learning [2, 32]. For selection, the feature vector from the
audio adaptive encoder \( A(X) \) is first processed by a fully
connected layer to give the activity probabilities \( h \in \mathbb{R}^K \).
Then, an initial vector is obtained by \( \mathbf{g} = \sum_{k=1}^K h_k \ast g_k \). We
include visual features to help silent activities select the rep-
resentative 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
\( \mathbf{z}_{\text{cls}} = \sum_{k=1}^K h_k' \ast g_k \), which serves as the class token.
Consequently, the input sequence for the transformer becomes:
\[
\mathbf{z} = [\mathbf{z}_{\text{cls}}; \mathbf{f}_{w, 1} \mathbf{E}_v + \mathbf{E}^d; \ldots; \mathbf{f}_{w, n} \mathbf{E}_v + \mathbf{E}^d],
\]
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 be-
comes 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 domains with the loss:
\[
\mathcal{L} = \sum_{(X, y_i) \in \{S, T\}} \mathcal{L}(\mathbf{p}_i^*, y_i) + \eta(\mathcal{L}(\mathbf{h}_i, y_i) + \mathcal{L}(\mathbf{h}'_i, y_i)),
\]
where hyperparameter \( \eta \) 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 par-
titions (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 dif-
f erent 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 in-
herent 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 adapta-
tion 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

### Table 1. Domain adaptation benchmarks for activity recognition under scenery, viewpoint and actor shift

<table>
<thead>
<tr>
<th>Shift</th>
<th>Video Dataset</th>
<th>Source Domain Setting</th>
<th>Target Domain Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Source Domain</td>
<td>Target Domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Train</td>
<td>Train</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test</td>
<td>Test</td>
</tr>
<tr>
<td>Viewpoint</td>
<td>CharadesEgo [33]</td>
<td>Third-person view</td>
<td>Ego-centric view</td>
</tr>
<tr>
<td>Actor</td>
<td>ActorShift (ours)</td>
<td>Human actors</td>
<td>Animal actors</td>
</tr>
</tbody>
</table>

**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.
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 \(r=3\), \(\gamma=0.05\) and \(\beta=0.0999\). 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

| 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.

<table>
<thead>
<tr>
<th>Model components</th>
<th>EPIC-Kitchens</th>
<th>CharadesEgo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1: Audio-adaptive encoder</td>
<td>51.2</td>
<td>23.1</td>
</tr>
<tr>
<td>+ Visual encoder (\mathcal{V}(\cdot))</td>
<td>48.0</td>
<td>23.1</td>
</tr>
<tr>
<td>+ Audio-based attention (\psi(\cdot))</td>
<td>51.2</td>
<td>23.5</td>
</tr>
<tr>
<td>+ Absent-activity learning</td>
<td>53.7</td>
<td>24.4</td>
</tr>
<tr>
<td>+ Audio-balanced learning</td>
<td>55.7</td>
<td>25.0</td>
</tr>
<tr>
<td>Stage 2: Audio-infused recognizer</td>
<td>55.7</td>
<td>25.0</td>
</tr>
<tr>
<td>+ Vanilla multi-modal transformer (\mathcal{M}^v)</td>
<td>56.1</td>
<td>25.0</td>
</tr>
<tr>
<td>+ Domain embedding (\mathcal{Z}^d)</td>
<td>57.2</td>
<td>25.4</td>
</tr>
<tr>
<td>+ Audio-adaptive class token (\mathcal{Z})</td>
<td>59.2</td>
<td>26.3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model components</th>
<th>Silent</th>
<th>Audible</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual encoder (\mathcal{V}(\cdot))</td>
<td>23.2</td>
<td>22.7</td>
<td>23.1</td>
</tr>
<tr>
<td>Full model</td>
<td>26.3</td>
<td>25.9</td>
<td>26.3</td>
</tr>
</tbody>
</table>

Stage 2: Audio-infused recognizer. For the audio-infused recognizer, we first consider a vanilla transformer. It takes as input \(z^m\) (Eq. 5), i.e., the audio and visual features from the audio-adaptive encoder, mapped by \(\mathcal{E}_v\) and \(\mathcal{E}_a\) into a common space, alongside a learnable class token. This only gives a marginal improvement in results. Adding the domain embedding \(\mathcal{E}^d\) to reduce domain-specific visual features in \(z^v\) (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 \(z\) (Eq. 7), we observe further improvements of 2.0% and 0.6%. 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 \(\sim3\)% 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.
We first note that our approach gives changes across domains, the class-balanced loss 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality</th>
<th>EPIC-Kitchen Activity Recognition Across Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RGB</td>
<td>Flow</td>
</tr>
<tr>
<td>I3D backbone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source-only [27]</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Munro and Damen [27]</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Planamente et al. [30]†</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Yang et al. [48]†</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kim et al. [22]</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Song et al. [34]</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>This paper</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SlowFast backbone</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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.

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-
We conclude our approach does not require paired training. Actor shift. We use unpaired videos, the performance remains unchanged. Viewpoint shift. videos to be robust to viewpoint shift. viewpoints as in previous methods. is both a large shift in appearance and a significant difference to both. All models struggle with silent activities when there is both a large shift in appearance and a significant difference by Munro and Damen, following the semi-supervised setting of Choi et al. 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. with the same I3D RGB backbone, for all amounts of labeled target videos. When adopting the SlowFast RGB backbone, we again further improve performance for all settings. In the supplementary, we also provide a favorable comparison with Li et al. 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 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, 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 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


