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

A. Implementation Details

Visual encoder. On all datasets, we adopt the mmaction2 [10] toolkit. The SlowFast [15] network is used to extract features from RGB frames for experiments on the EPIC-Kitchens, CharadesEgo and ActorShift datasets. We also extract features from the optical flow modality on EPIC-Kitchens by the slow-only network [15]. The networks are initialized with Kinetics pre-trained model weights. Such pre-training on a large dataset is common in domain adaptation, *e.g.*, for images (ImageNet) [7, 17, 40, 41] and videos (Sports-1M [20], Kinetics [6, 9, 22, 27, 34]).

Audio encoder. We adopt a ResNet-18 [19] for all datasets and initialize the weights from the VGGSound pre-trained checkpoint [4]. The last residual block and final classification layer are further trained on each dataset before generating pseudo-absent labels and audiovisual fusion as detailed in Section 3.

Attention module. The attention module consists of eight transformer encoder layers [13, 14, 42] and the parameters are randomly initialized. The inputs are intermediate audio features from the audio encoder (conv3 of the ResNet-18 network). The output of the class token passes through one fully connected layer to obtain the attention vector for the visual encoder.

Audio-infused recognizer. We use three transformer encoder layers with the same architecture as in [14]. The parameters are also randomly initialized. The sequence dimension D is set to 512 and each layer has 8 self-attention heads.

Training objective. On EPIC-Kitchens, We use a standard softmax cross-entropy loss as $\mathcal{L}$ in Eq. 2 and Eq. 8. Since CharadesEgo aims for multi-label classification, the sigmoid cross-entropy loss is adopted as the $\mathcal{L}$ in Eq. 2 and Eq. 8.

Inference. When using a single modality, the output from the activity recognizer $\mathcal{R}(\cdot)$ is directly used as the final recognition prediction. On EPIC-Kitchens [27], when using both RGB and optical flow, we average the predictions from the two modalities as the final classification result, following prior works [22, 27, 34].

*Currently at Terminus Group, China.

Labels for Target Domain	EPIC-Kitchens Top-1 (%) ↑	CharadesEgo mAP (%) ↑
Visual-based hard pseudo labels	51.7	23.9
Visual-based pseudo-absent labels	53.8	24.3
Audio-based hard pseudo labels	47.6	22.8
Audio-based pseudo-absent labels	55.7	25.0

Table 5. **Ablation of pseudo-absent labels.** Using the pseudo-absent labels predicted by audio for absent-activity learning is more effective than the visual counterpart. Hard pseudo labels from either modality results in inferior performance.

B. Audible Activities on CharadesEgo

The 13 audible activities we select for the ablation in Table 3 are: *someone is laughing*, *someone is cooking something*, *laughing at television*, *closing a door*, *talking on a phone/camera*, *closing a window*, *closing a refrigerator*, *washing some clothes*, *watching television*, *washing their hands*, *opening a window*, *opening a door*, *someone is sneezing*.

C. Ablation of Pseudo-Absent Labels

Audio *vs.* visual pseudo labels. We introduce absent-activity learning in Section 3 to increase the discriminability of our audio-adaptive encoder in the target domain. The pseudo-absent labels are obtained from the audio encoder pre-trained on the source domain. We validate the effectiveness of this setting in Table 5. Here, we consider three alternatives. First, using a pre-trained visual encoder on the source domain to get "Visual-based pseudo-absent labels" or "Visual-based hard pseudo labels". In the latter a one-hot pseudo label for each video is obtained by taking the class with the highest probability. We can also create "Audio-based hard pseudo labels" in the same way from the pre-trained audio encoder. Note that when using the hard pseudo labels, we adopt a standard classification loss, *i.e.* softmax cross-entropy or sigmoid cross-entropy loss, for unlabeled target domain data, instead of the loss for absent-activity learning.

Both visual-based pseudo-absent and hard pseudo labels result in inferior performance, since the visual activity appearance has larger variance across domains than audio and

Value of r	Top-1 (%) ↑
1	52.3
2	54.3
3	55.7
4	55.2
5	54.1
6	51.9

Table 6. **Effect of r** for audio-based pseudo-absent labels on single-label classification with EPIC-Kitchens. While a small r provides little supervision in the target domain, a large r also degrades the performance due to the unreliable predictions for silent activities. r=3 results in the best performance.

Value of γ	mAP (%) ↑
0.03	24.5
0.04	24.8
0.05	25.0
0.06	24.7
0.08	24.5
0.1	24.1

Table 7. **Effect of γ** for audio-based pseudo-absent labels on multi-label classification with CharadesEgo. γ=0.05 results in the best performance.

the visual encoder suffers from distribution shift. Since the audio predictions are unreliable for silent activities, using the hard pseudo labels from audio is worse than visual-based pseudo-labels. By contrast, our audio-based pseudo-absent labels provide reliable supervisory signals for the visual encoder adapting to the domain shift.

Effect of r. When we generate pseudo-absent labels for single-label classification on EPIC-Kitchens, r classes with the lowest audio-based probabilities are treated as the absent activities to train our audio-adaptive encoder. In all other experiments r=3 is used. We illustrate the effect of r in Table 6. When r equals 1, little supervision is provided for training the visual encoder in the target domain so that its adaptation ability degrades. With a large r, the pseudo-absent labels are noisy, since the audio predictions for silent activities are unreliable. Overall, we consider r=3 to be the best trade-off.

Effect of γ. For the multi-label classification, we assume the $(1-\alpha_k)\gamma$ percent videos with the lowest audio-based probabilities do not contain class k, where α_k is the percentage of videos containing class k in the labeled source domain. Then, we obtain the pseudo-absent label for each video according to this rule. Here, we study the effect of γ in Table 7. Similar to the effect of r, γ=0.05 delivers the best result, while a smaller or larger γ causes the performance to degrade slightly.

Audio variance across domains. The audio modality is more domain-invariant for distinguishing true negatives. When predicting absent activities in the source domain, both audio and visual classifiers achieve a high true negative rate on EPIC-Kitchens of 96.1% and 97.2%. In the target domain the audio remains robust with 95.6% true negative rate, while that of the visual classifier degrades to 86.7%.

	EPIC-Kitchens	**CharadesEgo**
Number of Clusters	Top-1 (%) ↑	mAP (%) ↑
4	52.9	23.5
5	53.4	23.9
6	53.8	24.2
7	54.3	24.5
8	53.9	24.1
9	53.6	23.8
10	52.7	23.3
Elbow method [37]	55.7	25.0

Table 8. **Effect of Elbow method** on our audio-adaptive encoder. Using a fixed number of clusters results in inferior performance compared to the Elbow method.

D. Ablation of Audio-Balanced Learning

For our audio-balanced learning, we use audio features to cluster the samples inside each class in the source domain, and each cluster is treated as one type of interaction with objects or environments.

Effect of Elbow method. The Elbow method [37] is adopted for determining the number of clusters for each class. It gives 5 to 12 clusters per class on both EPIC-Kitchens and CharadesEgo. Here, we compare its performance with a fixed number of clusters for all classes in Table 8. The Elbow method [37] results in the best performance. The reason is that some activity classes do not have large variance in the interactions, *e.g.*, *pouring*. For such an activity class, when using a large number of clusters, some clusters may contain samples with similar activity appearance to those in some other clusters of this class. Then, with audio-balanced learning, the visual encoder may pay more attention to these 'redundant' clusters during training, as they contain less samples, and over-fit to the samples in these clusters. Besides, if all the classes adopt a small number of clusters, some rare interactions with objects or environments in the source domain cannot be well learned and thus the overall accuracy of the model degrades on the target domain.

Audio *vs.* visual features for clustering. As an alternative to our audio-based clustering in Section 3, we could instead rely on visual features for clustering. However, since visual features are sensitive to appearance changes, such as the background color, we observe some of the resulting clusters may mainly contain videos with similar backgrounds, rather than a specific type of interaction. We compare the performance per modality in Table 9. Clustering by audio outperforms the visual counterpart by 2.3% top-1 accuracy

Modality	EPIC-Kitchens Top-1 (%) ↑	CharadesEgo mAP (%) ↑
Visual	53.4	24.1
Audio	55.7	25.0

Table 9. **Audio *vs.* visual features for clustering** in the audio-balanced learning for our audio-adaptive encoder. Using audio features is preferred over the visual features and delivers 2.3% top-1 accuracy and 0.9% mAP advantage.

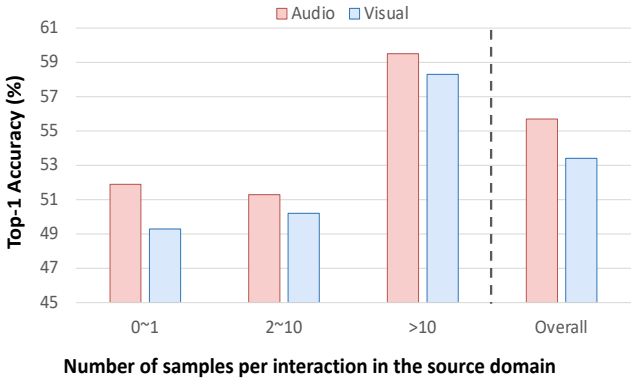

Figure 7. **Audio *vs.* visual features for long-tail** on EPIC-Kitchens. Clustering by audio features to identify rare and frequent activities for balanced learning is preferred over visual features under semantic distribution shift.

on EPIC-Kitchens and 0.9% mAP on CharadesEgo.

We also show the difference in performance towards rare activities in EPIC-Kitchens when using visual-based clustering in Figure 7. Using audio for clustering delivers higher accuracy towards rare activities than the visual counterpart. We also observe that clustering by the audio modality is better than clustering by visual features on frequent activities. This is because the model with visual clustering may focus more on rare differences in backgrounds that have no effect on the activity class. We conclude that forcing the model to learn in a balanced way towards different types of interactions clustered by the audio can better handle semantic distribution shift than the visual counterpart.

Visualizations of clusters. Some examples from the clusters are provided in the supplementary video. We observe that each cluster by the audio features tends to contain similar objects or environments the actors interact with. For example, for the *putting* activity in EPIC-Kitchens, one cluster has a bias towards food, one consists mainly of plates and another one prefers kitchen utensils. Similar phenomena also exist in CharadesEgo. For the *eating* activity, the actors are commonly watching television in the living room when eating in one cluster. Kitchens or bedrooms with silent environments are frequent sites for eating in another cluster. There is also one cluster in which there are several people chatting in each video while the actor of interest is eating.

Source of class token	EPIC-Kitchens Top-1 (%) ↑	CharadesEgo mAP (%) ↑
Audio classification prediction	57.5	25.6
Audio features	57.8	25.6
Audio features†	55.9	25.2
Activity sound feature vectors	59.2	26.3

† Only training the audio-infused activity recognizer.

Table 10. **Audio-adaptive class token ablation.** Obtaining the audio-adaptive class token from activity sound feature vectors is preferred over using the original audio feature vector or classification prediction from the audio encoder as the class token.

We conclude that when finding rare activities by clustering, using audio features is reliable.

E. Ablation of Audio-Infused Recognizer

Audio-adaptive class token. In Section 3, we obtain the audio-adaptive class token from a series of activity sound feature vectors, considering both audio and visual features. Alternatively, we can directly conduct global average pooling on the audio features from the audio encoder and treat the resulting feature vector as the class token. Then, the audio-infused activity recognizer can be jointly trained with the last residual block of the audio encoder. The performance comparison is shown in Table 10. Since joint training includes more parameters than training the recognizer only, the model suffers from over-fitting [44] and results in 1.4% top-1 accuracy and 0.7% mAP drops, compared to using activity sound feature vectors. However, if we fix the audio encoder and only train the recognizer, the performance degrades dramatically. This is expected as audio and visual features come from different feature distributions and cannot ensure effective audio-visual interaction by the attention inside the recognizer. We conclude our audio-adaptive class token from the activity sound feature vectors is superior compared to the original audio features from the audio encoder.

Recognizer *vs.* simple classifier. To further justify our audio-infused recognizer, we test an alternative classifier that uses the audio-attention weights h concatenated with the visual feature as input to a fully connected layer. It scores 55.9% top-1 accuracy on EPIC-Kitchens and 24.8% mAP on CharadesEgo. Worse than our 59.2% and 26.3%.

F. Effect of Depth in Transformer Modules

Attention module. Our attention module consists of eight transformer encoder layers. With more layers, the module may suffer from over-fitting. By contrast, only using a few layers will lead to under-fitting. We study the effect of depth on our audio-adaptive encoder by setting it in the range of $[5, 12]$, and the results on EPIC-Kitchens with RGB and

Method	Modality			EPIC-Kitchen Activity Recognition Across Domains						
	RGB	Flow	Audio	D2 → D1	D3 → D1	D1 → D2	D3 → D2	D1 → D3	D2 → D3	*Mean*
This paper			✓	37.2	38.4	40.7	44.7	46.0	47.0	42.3
I3D Architecture										
Munro and Damen [27]	✓	✓		48.2	50.9	49.5	56.1	44.1	52.7	50.3
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		✓		44.5	44.7	50.8	55.3	42.1	50.2	47.9
This paper		✓	✓	48.7	48.3	52.3	60.9	49.2	53.1	52.1
This paper	✓			38.1	37.8	39.2	47.9	42.1	41.8	41.1
This paper	✓	✓		43.7	42.6	45.7	50.2	43.8	52.3	46.4
This paper	✓		✓	49.5	43.5	49.2	54.6	46.6	50.0	48.9
This paper	✓	✓	✓	**51.9**	48.7	**53.2**	**63.2**	**52.1**	**55.5**	**54.1**
SlowFast Architecture										
This paper		✓		48.5	49.2	50.4	54.1	44.1	52.8	49.8
This paper	✓			52.4	51.1	51.8	55.9	45.7	53.6	51.8
This paper		✓	✓	50.8	51.5	52.7	60.7	48.4	57.7	53.6
This paper	✓	✓		53.9	54.3	53.5	58.7	47.9	55.2	53.9
This paper	✓		✓	57.9	55.9	58.4	64.3	54.2	**64.3**	59.2
This paper	✓	✓	✓	**59.3**	**59.1**	**59.5**	**69.1**	**54.8**	**64.3**	**61.0**

Table 13. **Modality ablation under scenery shift** on EPIC-Kitchens for the unsupervised domain adaptation setting. Relying on either audio or visual modality results in inferior performance, while our audio-adaptive models achieve state-of-the-art accuracy.

Depth	Top-1 (%) ↑
5	55.0
6	55.3
7	55.3
8	55.7
9	55.5
10	55.3
11	55.2
12	54.9

Table 11. **Effect of depth** for the audio-adaptive encoder on EPIC-Kitchens. The performance is not very sensitive to the depth and using 8 layers results in the best performance.

Depth	Top-1 (%) ↑
1	56.8
2	57.3
3	59.2
4	59.0
5	58.4
6	58.1

Table 12. **Effect of depth** for the audio-infused recognizer on EPIC-Kitchens. Increasing depth to 3 layers is effective, then performance plateaus.

audio modalities are shown in Table 11. Performance is not very sensitive to the depth, using 8 layers represents the best empirical trade-off.

Audio-infused recognizer. Our audio-infused recognizer contains 3 transformer layers. Similar to the attention module, more layers lead to over-fitting while less layers result in under-fitting. We study the effect of depth on EPIC-Kitchens with RGB and audio modalities and the results are shown in Table 12. Increasing the depth to 3 layers is effective, then performance plateaus.

G. Modality Ablation on EPIC-Kitchens

We provide more modality-combinations in Table 13. The audio encoder alone can achieve only 42.3% top-1 accuracy, since it cannot predict accurately on silent activities. We also consider using the visual modalities only, *i.e*. RGB and optical flow, and modify our approach correspondingly. To be specific, the pseudo-absent labels are determined by the pre-trained visual encoder in the source domain. Visual features are used for clustering in the audio-balanced learning as well as generating the attention for themselves. The activity recognizer also takes visual features only as inputs along with a learnable class token as in [14]. The visual-only versions of our approach achieve inferior performance since the activity appearance suffers from large variances under domain shift. However, our audio-adaptive model with either RGB or optical flow modality delivers competitive accuracy with the aid of the domain-invariant information within sound.

H. Additional Comparison on CharadesEgo

Li *et al*. recently proposed a supervised-only approach [26] for model pre-training to be better suited for downstream tasks with egocentric videos. We find our approach profits from their features as well. On its own the approach from Li *et al*. achieves 30.6 mAP on CharadesEgo. Combined with our audio-based attention and audio-infused recognizer, we improve this to 31.9 mAP.

Method	Top-1 (%) ↑
Sight or Sound	
Visual-only	48.0
Audio-only	42.3
Within-domain fusion	
Late fusion	47.2
Lee *et al*. [25]	50.8
Tian *et al*. [38]	50.0
Nagrani *et al*. [28]	51.1
Gabeur *et al*. [16]	51.3
Cross-domain fusion	
This paper	**59.2**

Table 14. **Fusion benefit.** Experiments performed on EPIC-Kitchens with the same RGB and audio modalities. Although utilizing within-domain fusion methods can achieve good performance, our approach provides more effective cross-modal interaction under domain shift.

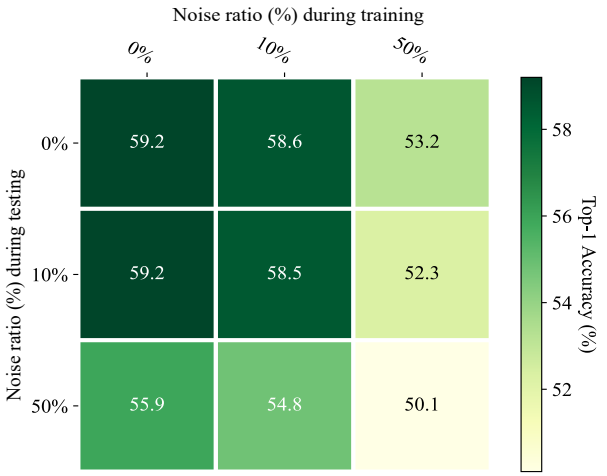

Figure 8. **Impact of audio quality** on EPIC-Kitchens. Our model remains robust up to 10% irrelevant sound during training and/or testing. With more noise, especially during training, performance starts to suffer.

Figure 9. **Example activities from the ActorShift dataset**.

I. Fusion Benefit

We also compare our full audio-adaptive model with alternative audio-visual fusion approaches originally intended for within-domain activity recognition [16, 25, 28, 38] on EPIC-Kitchens. We use either publicly available implementations or re-implement ourselves and let them all use the same inputs, *i.e*., the features as outputted by the visual and audio encoders. The results are shown in Table 14. We denote simple averaging of the classification predictions from both encoders by "Late fusion". As most activities are silent, the audio predictions are unreliable and degrade the performance when combined with visual predictions via late fusion. Although the cross-modal interaction methods proposed in [16, 25, 28, 38] are designed for within-domain activity recognition, they still achieve good performance compared to a visual-only encoder with 48.0% top-1 accuracy. This is expected as several samples in the target domain may not contain a large domain shift, so the audio-visual correspondences from the source domain will also be encountered in the target domain. Our full audio-adaptive model allows for an even more effective cross-modal interaction under domain shift. This is because our audio-adaptive encoder and audio-infused recognizer alleviate the dependence on searching cross-modal correspondences for classification and instead rely on the domain-invariant activity information within sound to obtain a more discriminative visual feature representation in the target domain.

J. Audio Quality Assumption

Throughout our work, we assume the audio track accompanying a video is of decent quality. To measure the impact of audio quality, we mix the audio track of each video with the audio from another randomly sampled video. In Figure 8 we vary the noise ratio in train and test and measure the top-1 accuracy (%) on EPIC-Kitchens. Our model remains robust up to 10% irrelevant sound during training and/or testing. With more noise, especially during training, performance starts to suffer.

K. Details of ActorShift Dataset

When constructing the dataset, we first select 7 Kinetics classes for which there are dozens of videos of animals performing these actions on YouTube. The videos are collected by querying YouTube with 'animals' prepended to the verb of the action class, *e.g.,* "animals sleeping". We discard videos with music and only keep those with the original animal sounds. Videos with the animal out-of-view are also rejected. The final dataset covers a wide range of animal species including dog, deer, koala, cat, alpaca, lion, tiger, kangaroo, loris, raccoon, rabbit, elephant, monkey, panda, horse, duck, bird, snail, cow, chinchilla, marmot, lizard, hedgehog, bat, tortoise, squirrel, giraffe, goose and fox. Some examples are shown in Figure 9.

L. Supplementary Video

We provide more visualizations in the supplementary video on our project page: https://xiaobai1217.github.io/DomainAdaptation. It includes examples about the pseudo-absent labels for absent-activity learning and the clusters generated by audio features for audio-balanced learning. We also compare the predictions between a visual-only encoder and our audio-adaptive encoder, as well as the benefit brought by our audio-infused recognizer. Some failure cases on the ActorShift dataset are also shown, where the domain shift exists in both the visual and audio modalities.

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.