EPIC-SOUNDS: A LARGE-SCALE DATASET OF ACTIONS THAT SOUND

Jaesung Huh[1*], *Jacob Chalk*[2*], *Evangelos Kazakos*[2†], *Dima Damen*[2], *Andrew Zisserman*[1]

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford, UK
[2]Department of Computer Science, University of Bristol, UK

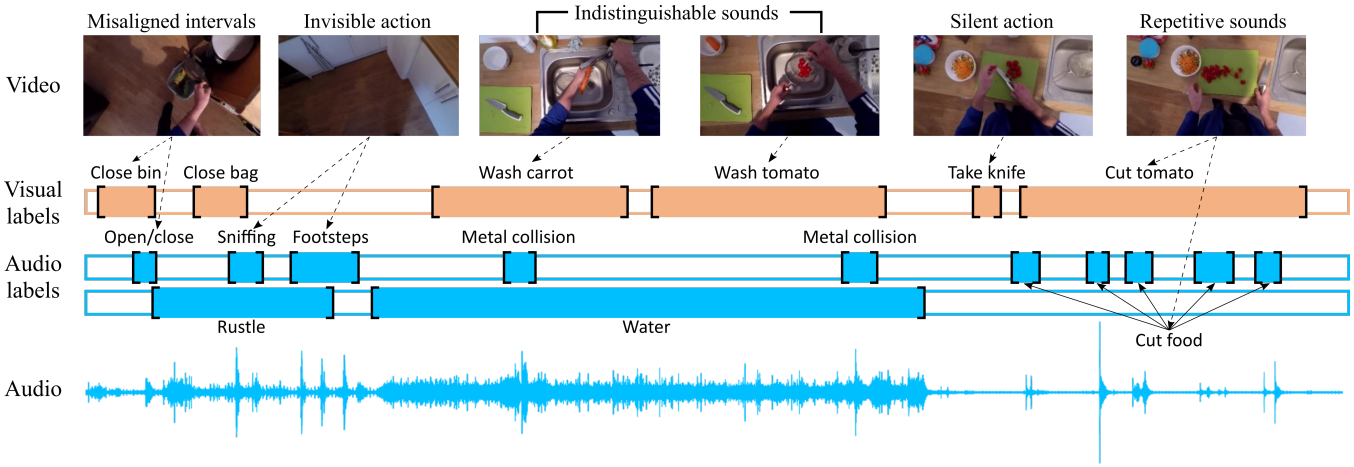

Fig. 1: Sample video with corresponding audio from EPIC-KITCHENS-100 [1]. We compare the already published **visual labels** with our collected EPIC-SOUNDS **audio labels**. We demonstrate the differences between the modality annotations, both in temporal extent and class labels, highlighting: **Misaligned intervals:** temporal boundaries are distinct; **Invisible action:** action not seen in the video, but which produces distinct sounds (0-to-1 matching); **Indistinguishable sounds:** sounds from two distinct visual actions, but are audibly inseparable; **Silent action:** visual action that does not have audible sounds (1-to-0); and visual actions containing multiple **repetitive sounds** (1-to-N).

ABSTRACT

We introduce EPIC-SOUNDS, a large-scale dataset of audio annotations capturing temporal extents and class labels within the audio stream of the egocentric videos. We propose an annotation pipeline where annotators temporally label distinguishable audio segments and describe the action that could have caused this sound. We identify actions that can be discriminated purely from audio, through grouping these free-form descriptions of audio into classes. For actions that involve objects colliding, we collect human annotations of the materials of these objects (e.g. a glass object being placed on a wooden surface), which we verify from visual labels, discarding ambiguities. Overall, EPIC-SOUNDS includes 78.4k categorised segments of audible events and actions, distributed across 44 classes as well as 39.2k non-categorised segments. We train and evaluate two state-of-the-art audio recognition models on our dataset, highlighting the importance of audio-only labels and the limitations of current models to recognise *actions that sound*.

Index Terms— audio recognition, action recognition, audio event detection, audio dataset, data collection, dataset

1. INTRODUCTION

Humans perceive objects and actions through multiple senses, especially vision and audition [2]. Inspired by this, a plethora of works aim to solve various video understanding tasks, such as action recognition [3, 4, 5] and detection [6, 7], by fusing the two modalities. These attempts are especially common for egocentric video datasets due to the camera's close proximity to the ongoing actions resulting in clearer inputs, both visually and audibly. Research has shown improved performance by using audio and video jointly in egocentric data [8, 9, 10, 11].

In general, these works make two key incorrect assumptions: First, that the visual and auditory events temporally coincide; Second, that a single set of classes can be used for both modalities, typically derived from vision. In practice, visual and auditory events exhibit varied levels of both temporal and semantic congruence, thus violating these assumptions (See Figure 1). In the case of actions such as 'close bin', the onset of the visual event can be defined as the time that the person grasps the handle, whereas the onset of the audio event is delayed to the moment when the lid of the bin slams. Some actions are audibly indistinguishable, e.g. 'wash carrot' vs 'wash tomato', as it is impossible to determine which vegetable is being washed through sound alone. Consequently, using the visual temporal labels as targets for training an audio classifier is often a flawed endeavour – the resulting audio classifier will not be able to discriminate all of the visual events; and many audio labels that could provide supervision for training are missed. Based on these observations, we crowdsource temporal and semantic labels for the audio of EPIC-KITCHENS-100 that are distinct from the visual ones.

However, as evidence suggests [12], humans perform poorly at recognising objects and events using audio alone, making their annotation using only audio challenging. Due to the lack of sufficient information in audio for inferring fine-grained properties of events, humans tend to use vague terms for describing them; *e.g.* when the interaction from the collision of two objects is indistinguishable from audio, annotators often describe the associated event as 'clang' or 'bang'. To alleviate this, we further augment these semantics with

*Equal technical contribution. † Now at Samsung AI Center Cambridge.

Table 1: Comparison to existing datasets. **A**: Audio. **V**: Video. **T**: Temporal annotations. We showcase that EPIC-SOUNDS is the only dataset with distinct classes for audio and video modalities (**D**). We only report categorised segments of EPIC-SOUNDS here.

Name	Source	# hrs	# seg.	# cls	Modality	T	D
DESED [13]	real + synth.	43h	8k	10	**A**	✓	N/A
URBAN-SED [14]	synth.	30h	50k	10	**A**	✓	N/A
TUT 2016 [15]	real	2h	6.3k	18	**A**	✓	N/A
AudioSet [16]	YouTube	5833h	1.8M	632	**A + V**	✗	✗
VGG-Sound [17]	YouTube	550h	200k	309	**A + V**	✗	✗
SSW60 [18]	real	25.7h	9.2k	60	**A + V**	✗	✗
LLP [19]	YouTube	33h	19.4k	25	**A + V**	✓	✗
EPIC-SOUNDS	home kitchens	100h	78.4k	44	**A + V**	✓	✓

the *materials* of the objects that interact. We verify these from the video, discarding incorrect audio-only material annotations.

In summary, we introduce EPIC-SOUNDS, a large-scale dataset of daily-life sounds, derived from the audio of EPIC-KITCHENS-100. EPIC-SOUNDS contains 78,366 categorised sound events spanning over 44 categories, as well as 39,187 non-categorised sound events, totalling 117,553 sound events across 100 hours of footage collected in 700 videos from 45 home kitchens. The sound classes are based on descriptions from only listening to audio, thus suitable for problems in acoustics such as audio/sound recognition and sound event detection. EPIC-SOUNDS is available from: https://epic-kitchens.github.io/epic-sounds.

2. RELATED WORK

Sound event detection datasets. Sound Event Detection (SED) is the task of detecting the onset and offset of audio events as well as recognising the event within the detected boundaries. SED datasets [13, 14, 15] are similar to EPIC-SOUNDS as these include annotations of temporal boundaries of events, whereas sound recognition datasets [20, 21, 22] do not. Nevertheless, they differ from EPIC-SOUNDS in several aspects. First, they are of smaller scale making the training of modern architectures impractical. Second, [13] and [14] contain synthetic audio, and therefore models trained on these datasets generalise poorly to real recordings. Third, [13, 14, 15] contain sounds associated with generic scenes and events, whereas EPIC-SOUNDS focuses on fine-grained sounds generated from diverse audible events in 45 home kitchens.

Audio-visual datasets. We compare EPIC-SOUNDS to publicly available sound recognition or detection datasets in Table 1. AudioSet [16] is the largest audio-visual dataset of audio events with 2.1M clips and 527 annotated classes, while VGG-Sound [17] contains over 200K video clips and 300 audio classes. They are both collected from YouTube and each audio clip is 10s long. Both do not have temporal annotations for events, and importantly, a single set of annotations is collected for both modalities. The LLP dataset [19] is the closest to ours, in that both visual and auditory events are annotated independently, providing separate temporal segments. However, unlike ours, both modalities still share the same label set. Also, LLP is of smaller scale and contains diverse events while EPIC-SOUNDS focuses on sounds resulting from actions.

Fine-grained audio-visual datasets. The PACS dataset [23] focuses on understanding the physical common sense attributes of objects shown in the video, which is similar to our 'material' based annotation procedure. However, these attributes are distinguished by 13.4K question-answer pairs; displaying the video with and without audio, and then querying a variety of physical properties. SSW60 [18] consists of 31K images, 3.8K audio and 5.4K videos of 60 species of birds, proposed to facilitate works on fine-grained categorization using audio-visual fusion. Both datasets do not contain temporal annotations of sounds.

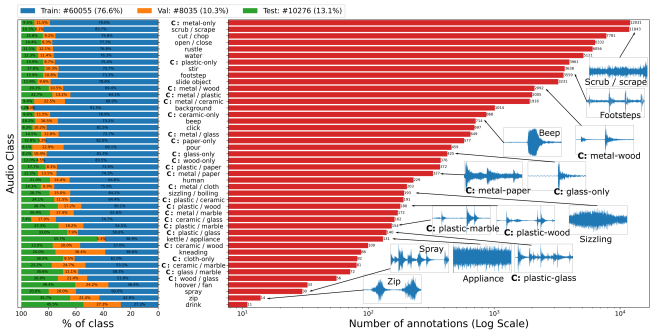

Fig. 2: Left: The percentage distribution of each audio class across the EPIC-SOUNDS dataset splits. Right: Class frequencies showcasing the long-tail distribution. **C:** represents a collision-based sound between objects of the same or two distinct material types.

3. EPIC-SOUNDS: DATASET STATISTICS

EPIC-KITCHENS-100. EPIC-KITCHENS-100 [1] is a large-scale egocentric audio-visual dataset which contains 100 hours of videos containing unscripted daily activities and object interactions in people's kitchens. It consists of 700 videos and 89,977 segments describing visual actions that occur. Actions consist of verb and noun labels, where there are 97 verb classes and 300 noun classes. The average action length is 2.6s. Since these actions are based only on video, we emphasise that we do not refer to any of these labels during the annotation process.

EPIC-SOUNDS. The dataset consists of 78,366 categorised temporal annotations with an average length of 4.9s, distributed across 44 classes. We match the train / val / test splits from EPIC-KITCHENS-100, giving the per-class proportion across splits in Fig. 2 (left). We divide the test split into two roughly even subsets: one for audio-based interaction recognition, and one for audio-based interaction detection. We release start/end times for the recognition subset, and keep those for the detection hidden for the relevant challenge.

Class frequency is also shown in Figure 2 (right), highlighting that EPIC-SOUNDS is naturally long-tailed. We also visualise the waveforms for a sampled subset of the classes. Here, there are both classes which produce waveforms consistent with short-term, percussive sounds such as all the collision-based classes, as well as long-term sounds e.g. sizzling. We also visualise the length of the annotations distributed across the classes in Figure 3. Here, we sort the classes by the median of their lengths, $\tilde{t}$, and distinguish three categories: long-term ($\tilde{t} \geq 10s$); intermediate ($1s < \tilde{t} < 10s$); and short-term ($\tilde{t} \leq 1s$) classes. Long-term classes relate to lengthier activities, such as cooking and hoovering. In the intermediate classes, there are sounds such as scrub / scrape, or rustle, and then near instantaneous/percussive sounds in the short-term category, including all collision-based classes.

4. DATA COLLECTION PIPELINE

The data collection process is conducted through the collection of temporal segments of distinct sounds, described by free-form vocabulary, followed by clustering generic sound categories into distinct classes. This section details this process, as well as post-processing steps taken to refine the results.

4.1. Data collection of labelled temporal segments

The objective is to annotate all the distinctive audio events that occur across all the videos in EPIC-KITCHENS-100. The annotation consists of the temporal interval of the event, together with a free-form text description. As the video length in this dataset varies greatly,

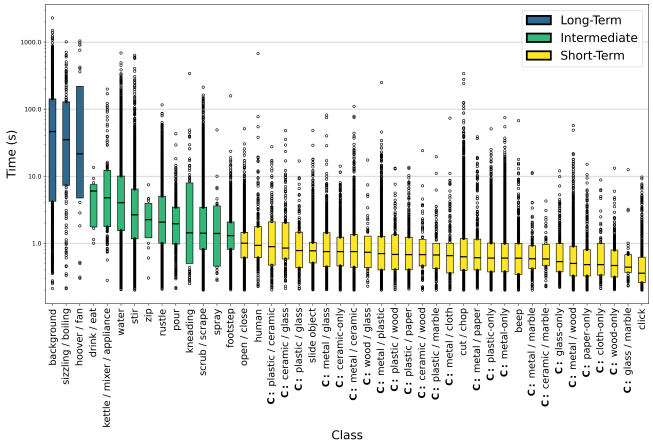

Fig. 3: Box plot for the lengths of the annotations over classes, ordered by the median of their lengths. The majority of the classes, 30 (68%) are short-term, 11 (25%) are intermediate classes and only 3 (7%) are considered long-term (median $>$ 10s). **C:** collision-based sounds between objects of the same or two distinct material types.

from 30 seconds to 1.5 hours, we trim the videos into a series of manageable lengths for annotations of 3-4 minutes. We deem our decision to only provide the audio stream as a key step so the annotators focus on the temporal bounds of the acoustic event alone, rather than being biased by visual and contextual information present in the video stream (consider the 'misaligned intervals' example shown in Figure 1, where visual and auditory temporal segments do not align for the same event). However, the annotators are provided with the plotted audio waveform to act as a visual guide to assist in targeting specific audio signatures and streamline the annotation process.

Instructions to the annotators. We worked with 20 annotators hired from an annotation company. We use a customised version of the VIA tool [24] to gather the annotations. Annotators are asked to listen to the audio and detect any distinctive audio event. They then are instructed to mark the start and end time of each distinctive sound they hear. Each segment is then given a semantic label which best describes the annotator's perception of the action associated with the audio event. We impose no restriction on the vocabulary used, so the annotators may describe this however they wish. As a guide, we provide a list of sound labels that commonly occur in daily life, which the annotator may refer to, though they are not required to explicitly choose from this list. We term a segment-label pair as an 'audio annotation'. A second annotator performs quality assessment to the audio annotations produced by the first annotator, particularly focusing on any missed audio events.

For each unique label description, the VIA tool creates a separate time-line, effectively grouping sequences of the same event. Note that sound events can overlap in time. If two segments are less than 0.3s apart, we instruct the annotators to merge the two segments as we deem them to belong to the same event. Additionally, annotators are asked to identify consistent background sounds (or noise) that occur throughout a large portion of the audio (e.g. radio, fan or washing machine). The annotators were asked to tag these as 'background'. The procedure described thus far resulted in the annotation of 556 distinct sound descriptions.

Humans tend to use abstract words to describe sounds, such as 'clang' or 'clatter', especially for those generated from the collisions between objects. To address this, we use a customised LISA [25] annotation interface for annotating the material of the objects that collide based on audio. We instruct annotators to select from a pre-specified list which *materials* are involved in the collision. This list, along with examples of objects for each material, are provided in Table 2. These cover all the materials popular in kitchens. Annotators are encouraged to select one or more materials, or mark the material as indistinguishable by choosing the 'Can't tell' option. We drop the instances in the latter case – as we believe these are unhelpful for sound or event understanding tasks. However, some material sounds might be deceiving. For example, one might perceive the material collision to be between a glass and a wooden object, but in fact it's food poured into a ceramic container. We thus ask annotators to then visually verify their material annotations using the corresponding video. Importantly, annotators have to listen and choose the perceived material first, and cannot change these after watching the video. Instead, they select the actual materials involved when viewing the video. We only retain visually-verified collision sounds – i.e. materials correctly perceived from the audio only, then verified from the visual observation. We choose all collision material labels for which at least 40 examples are present. As a result, abstract labels related to collision (e.g. 'clang/clatter', 'put objects on surface') are clustered into 24 sound categories describing the materials involved, such as **C**: metal-only, or **C**: plastic-wood. We use the letter **C** to indicate these are collision-based classes.

Table 2: Material options for collision sounds. We note **# of time** each material was **selected** in collision sounds, and discard the sounds annotated with 'Others' or 'Can't tell'.

Material	Example objects	# of times selected
Metal	metal or stainless steel	15523
Plastic	plastic bowl, plastic container	5464
Ceramic	ceramic cup, plate	2634
Wood	wooden spatula, wooden table	2408
Paper	kitchen roll, cardboard boxes	1253
Glass	wine glasses, glass cup	1248
Stone / Marble	kitchen worktops, marble tables	377
Cloth	towels, teatowels, clothes	257
Others	materials not listed above (e.g. food)	3596
Can't tell	cannot determine the material	10030

4.2. Post-processing Annotations

From labels to classes. We post-process the audio labels to fix spelling errors and group semantic equivalences. For example, sounds like 'buzzer', 'beep' and 'alarm' are grouped into one *beep* class. Similarly, sounds described by the verbs 'wipe', 'scour', 'scrape' and 'scrub' are also grouped into a single class. We also manually review tail instances to determine whether these form novel classes or should be merged with others. In cases where the description was not meaningful, the categorised annotation is dropped. For example, the sound 'spray' was considered a meaningful tail instance of an action that sounds. In contrast, the label 'dog barking' was discarded as it is not relevant to our context. This produces the 44 audio classes, as shown in Figure 2.

Error checking audio classes. Due to differences in sound perception between annotators, some errors exist amongst the classes. For example, where one annotator hears a drawer being pulled and hence labels 'open / close', another may hear 'drag object' for a similar audio. To resolve such errors, we manually review each of the labels in the test and validation set. Specifically, the following procedures are conducted to correct samples in the validation and test set. (1) We ask annotators to manually review all the val / test samples, providing them only sounds for non-collision classes and sounds and corresponding video clips for collision classes. (2) We collect the

Table 3: Results of the Baseline Models on the EPIC-SOUNDS validation, recognition test and entire test splits. L: Linear-Probe; F: Fine-Tuning.

Split	Model		Top-1	Top-5	mCA	mAP	mAUC
Val	Chance	-	7.71	30.95	2.29	0.023	0.500
	SSAST [26]	L	28.74	64.87	7.14	0.079	0.755
	ASF [27]	L	45.53	79.33	13.48	0.172	0.789
	SSAST [26]	F	53.47	**84.56**	**20.22**	0.235	**0.879**
	ASF [27]	F	**53.75**	84.54	20.11	**0.254**	0.873
Recognition Test	Chance	-	7.85	31.91	2.39	0.024	0.500
	SSAST [26]	L	29.93	66.60	7.17	0.082	0.725
	ASF [27]	L	45.00	78.98	15.00	0.183	0.788
	SSAST [26]	F	53.71	84.54	**22.28**	0.223	0.820
	ASF [27]	F	**54.45**	**85.17**	20.41	**0.254**	**0.852**
Entire Test	Chance	-	7.22	30.11	2.27	0.023	0.500
	SSAST [26]	L	27.50	65.55	6.68	0.080	0.741
	ASF [27]	L	44.55	78.44	14.49	0.145	0.772
	SSAST [26]	F	53.75	83.76	**20.76**	**0.237**	**0.860**
	ASF [27]	F	**54.86**	**84.26**	20.30	0.232	0.823

samples in which the first and second annotations are inconsistent, and ask a new set of annotators to manually choose whether the 1st or 2nd annotations are correct. The annotators could choose 'can't tell' or 'neither of the two'. (3) We manually verify those decisions, and removed the samples from the val/test sets. For the training set, we utilise the overlaps between audio segments and visual segments to select the samples for reviewing. We deem the use of the visual labels acceptable for error correction, as the annotation process is complete. Thus, utilising the visual labels for post-processing no longer compromises the issues stated in Figure 1.

We review all audio classes for which there exists a *mapping* to visual classes in EPIC-KITCHENS-100. We identify two types of mapping, trivial; the audio class itself already exists as a visual class e.g. 'scrub', and relational; the audio class does not exist as a visual class itself but can be semantically mapped to one or more of the visual classes, such as the audio class 'click' relating to the verb 'turn on/off' or the noun 'light switch'. We consider all annotations *not* labelled as the audio class of interest, but which overlap with action clips containing its visual mappings. We then manually assess each overlapping annotation, through listening to the audio and determining the label's correctness. We run this error checking cycle multiple times to ensure all incorrectly classified instances are accounted for.

Non-categorised audio events. As a result of post-processing, there are audio events that we recognise the sound exists but no semantic label matching the 44 classes could be given. These are samples we either could not assign class labels, or collision sounds for which they could not be visually verified. We release these temporal boundaries of these 39,187 samples as *non-categorised*.

5. EXPERIMENTS AND RESULTS

This section describes how two state-of-the-art sound recognition models perform on classifying EPIC-SOUNDS. We assess models through performance metrics and class confusion matrices.

Baselines. We train and evaluate the Auditory SlowFast (ASF) [27] and Self-Supervised Audio Spectrogram Transformer (SSAST) [26] audio encoder networks, with both a linear probe, i.e. by freezing the model weights and only training the last classification layer, and by fine-tuning. We also compare to a chance baseline. ASF is pretrained on VGG-Sound, and SSAST is pretrained on AudioSet and LibriSpeech [28].

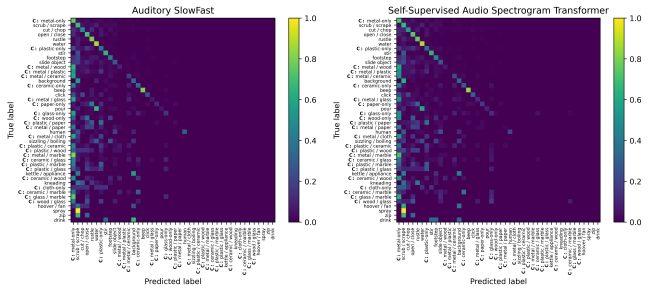

Fig. 4: Confusion Matrices on Val for ASF (left) and SSAST (right).

Audio processing. We follow the audio processing of [27] for extracting the input spectrograms for both models, noting that this outperformed the default audio processing of SSAST (200×128 spectrograms for 2s of audio, or 400×128 for 4s of audio sampled at 16kHz). Namely, audio is resampled at 24kHz for both models. We randomly sample 2s of audio to create log-mel-spectrograms with 128 Mel bands. If the audio annotation is shorter than 2s we pad the produced spectrogram with its last column. We use a window and hop size of 10ms and 5ms respectively, resulting to a spectrogram of size 400×128.

Training & Validation Configuration. We train both models for 30 epochs, setting the initial learning rate to $1\mathrm{e}{-3}$ for ASF which decays to 10% on epoch 25 and $1\mathrm{e}{-4}$ for SSAST, which is warmed up from $1\mathrm{e}{-6}$ for 2 epochs and decays to 5% then 1% on epochs 10 and 20. Both models are trained with cross-entropy loss, optimising ASF using SGD with Nesterov momentum equal to 0.9, and SSAST using AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$. Both models use a weight decay of 0.0001 and a batch size of 128. We use a base 384×384 ViT with patch size 16 as the backbone for SSAST and the 8×8 ResNet50 variant of ASF. For data augmentation, SpecAugment [29] is used, again following [27], using two frequency masks with $F = 27$, two time masks with $T = 25$ and time warp with $W = 5$. We use test augmentations simliar to [27], dividing the audio into 5 equally sized sub-clips and then averaging their individual predictions from the networks. For the linear probe results, we freeze the backbone of both SSAST and ASF and train only the last linear layer with the same training hyperparmeters and pretrained backbones as before.

Evaluation Metrics. We report the top-1 and top-5 accuracy, as well as mean average precision (mAP), mean area under ROC curve (mAUC), and mean per class accuracy (mCA), for both the validation and test sets.

Results. We report quantitative results for both models in Table 3. Overall, ASF outperforms SSAST by 0.28%, 0.74% and 1.11% for top-1 accuracy on the validation, recognition test and entire test set respectively. ASF exhibits better mAP for the validation set and recognition test set, whereas SSAST performs better on the entire test set, suggesting these models share a similar level of robustness to the long-tailed data. The performance of the linear probe drops significantly compared to fine-tuning results for ASF and almost halves for SSAST. In the latter case, we note that self-supervision alone does not learn class-discriminative features.

Figure 4 shows the validation confusion matrices for finetuned ASF and SSAST. We see that both models are able to detect a subset of distinctive, unique sounds such as rustle, water and beep. Concerning the collision-based classes, both models tend to classify unimaterial collisions more successfully than bi-material collisions, but generally produce a false positive prediction of the metal-only collision class, suggesting that the models may struggle to detect how

material properties alter the sound produced from a collision.

Reflections. When comparing audio to video labels, we reflect on our motivation in Figure 1. The top-3 audio classes that have 1-to-0 overlap with visual classes are: wood / glass collision (51.78%), metal / marble collision (51.67%), glass / marble collision (51.39%). In this instance, the classes relate to sounds produced by visual actions such as placing objects which are occasionally deemed trivial, or happen off-screen, resulting in missed visual annotations while still producing distinctive auditory signals. The top-3 video classes that have no overlap with audio (0-to-1) are: take basket (71.43%), pour basil (71.43%) and brush oil (70.0%) – these are silent actions. The top-3 many-to-1 classes that contain repeated audio sounds (on average) are: cut / chop (1.77-to-1), beep (1.31-to-1), metal-only collision (1.18-to-1), these relate to actions that have a 'stop-start' pattern e.g. pauses between chops, button presses on an appliance, or between repetitively moving items in a cutlery draw or sink.

6. CONCLUSION

In this paper, we present a large-scale dataset, EPIC-SOUNDS, which consists of 78.4k categorised segments and 39.2k non-categorised segments, totalling 117.6k segments spanning 100 hours of audio, capturing diverse actions that sound in home kitchens. Sound categories are annotated based on audio human descriptions. We also provide benchmark classification performance using the state-of-the-art sound recognition networks. The audio annotations in this dataset enable a veridical evaluation of audio classifiers, and can replace the current evaluations based on visual annotations. We anticipate that multi-modal approaches will benefit from these audio labels. The dataset can also be used for audio event detection, though we have not evaluated that in this work. The dataset and baseline code will be made publicly available.

Acknowledgements. This work proposes a new dataset that is publicly available, and builds on publicly available dataset EPIC-KITCHENS. Research and is supported by EPSRC Doctoral Training Program, EPSRC UMPIRE (EP/T004991/1) and the EPSRC Programme Grant VisualAI (EP/T028572/1). Jaesung Huh is funded by a Global Korea Scholarship. This project acknowledges the use of the EPSRC funded Tier 2 facility, JADE-II. We also thank Rajan from Elancer and his team, for their huge assistance with annotation.

7. REFERENCES

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray, "Rescaling egocentric vision: Collection pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision (IJCV)*, 2021.

[2] Linda Smith and Michael Gasser, "The development of embodied cognition: Six lessons from babies," *Artificial life*, vol. 11, no. 1-2, pp. 13–29, 2005.

[3] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani, "Listen to look: Action recognition by previewing audio," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10457–10467.

[4] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman, "Speech2action: Cross-modal supervision for action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10317–10326.

[5] Arsha Nagrani, Shan Yang, Anurag Arnab, Cordelia Schmid, and Chen Sun, "Attention bottlenecks for multimodal fusion," in *NeurIPS*, 2021.

[6] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018, pp. 247–263.

[7] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla, "Hear me out: Fusional approaches for audio augmented temporal action localization," *arXiv preprint arXiv:2106.14118*, 2021.

[8] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[9] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "With a little help from my temporal context: Multimodal egocentric action recognition," in *British Machine Vision Conference (BMVC)*, 2021.

[10] Merey Ramazanova, Victor Escorcia, Fabian Caba Heilbron, Chen Zhao, and Bernard Ghanem, "Owl (observe, watch, listen): Localizing actions in egocentric video via audiovisual temporal context," *arXiv preprint arXiv:2202.04947*, 2022.

[11] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid, "M&m mix: A multimodal multiview transformer ensemble," *arXiv preprint arXiv:2206.09852*, 2022.

[12] Nancy Jean VanDerveer, *Ecological acoustics: Human perception of environmental sounds*, Cornell University, 1979.

[13] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[14] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[15] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[17] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A large-scale audio-visual dataset," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[18] Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie, "Exploring fine-grained audiovisual categorization with the ssw60 dataset," in *ECCV*, 2022.

[19] Yapeng Tian, Dingzeyu Li, and Chenliang Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *ECCV*, 2020.

[20] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018, ACM Press.

[21] Marc Moreaux, Michael Garcia Ortiz, Isabelle Ferrané, and Frédéric Lerasle, "Benchmark for kitchen20, a daily life dataset for audio-based human action recognition," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2019, pp. 1–6.

[22] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[23] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency, "Pacs: A dataset for physical audiovisual commonsense reasoning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[24] Abhishek Dutta and Andrew Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, MM '19, ACM.

[25] *VGG List Annotator (LISA)*, 2022, https://www.robots.ox.ac.uk/ vgg/software/lisa/.

[26] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 10699–10709.

[27] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Slow-fast auditory streams for audio recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[29] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*. sep 2019, ISCA.