Dense-Localizing Audio-Visual Events in Untrimmed Videos: A Large-Scale Benchmark and Baseline (Supplementary Material)

Tiantian Geng[1,2], Teng Wang[1,3], Jinming Duan[2], Runmin Cong[4], Feng Zheng[1,5*]
[1]Southern University of Science and Technology [2]University of Birmingham
[3]The University of Hong Kong [4]Shandong University [5]Peng Cheng Laboratory
gengtiantian97@gmail.com tengwang@connect.hku.hk j.duan@bham.ac.uk
rmcong@sdu.edu.cn f.zheng@ieee.org

1. More Statistical Analysis

Concurrent Events. There are usually multiple audio-visual events occurring simultaneously in UnAV-100 dataset as in real-life scenes. Here, we define the overlap rate $\mathcal{O}$ of each video as:

$$\mathcal{O} = \frac{U_o}{U_e}, \tag{1}$$

where U_o is the temporal union of overlapping intervals, and U_e is the temporal union of the intervals of all audio-visual events in the video. Totally, there are around 25% of videos (2,651) containing concurrent audio-visual events ($\mathcal{O} > 0.01$, considering annotation errors) in our UnAV-100 dataset. The overlap rate distribution of these videos is illustrated in Fig. 1. We can see that the videos with low and high overlap rates both have high proportions. Higher overlap rates might indicate that the events have higher correlations and usually occur at the same time, which requires the model to have a strong ability of dependency modeling.

Temporal Dependencies between Events. We show NPMI (Normalized Pointwise Mutual Information) [4] of the pairs of simultaneous and consecutive audio-visual events for all 100 event categories in Fig. 2(a) and Fig 2(b), respectively. NPMI is commonly used in linguistics to represent the co-occurrence between two words. Firstly, in Fig. 2(a), we can observe that the event categories from the same domains are more likely to occur concurrently, *e.g.*, the events of human activities, music performances, and the sounds of vehicles/natural. Besides, the events from various domains are usually accompanied by human activities, *e.g.*, *playing acoustic guitar* with *male singing*, *basketball bounce* with *people crowd*, *etc*. Secondly, in Fig 2(b), in addition to the NPMI for consecutive occurrences of different audio-visual events, we also compute the values for the events from the same categories, which might be larger than 1. It can be observed that the same events tend to occur repetitively in a video, especially for some events that usually happen in a short period of time, such as *people nose blowing*, *people sneezing* and *basketball bounce*, *etc*. Moreover, diverse consecutive dependencies also exist between different audio-visual events.

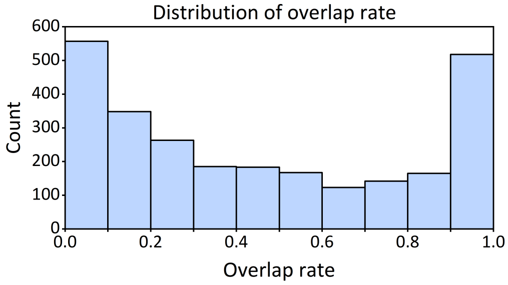


Figure 1. Overlap rate distribution of the videos that contain concurrent events in our UnAV-100 dataset.

Dataset	Videos	Classes	Avg. Length	Avg. Instances	Domains
Breakfast [7]	1,712	48	162s	6	Cooking
THUMOS14 [6]	413	20	212s	15.5	Sports
ActivityNet [1]	19,994	200	115s	1.5	Human Activities
Charades [8]	9,848	157	30s	6.8	Daily Activities
UnAV-100 (ours)	10,790	100	42s	2.8	Unconstrained

Table 1. Comparison with temporal action localization datasets based on untrimmed videos.

Comparison with Existing TAL Datasets. In Tab. 1, we compare our UnAV-100 dataset with four popular benchmarks for temporal action localization. All these datasets are based on untrimmed videos and have relatively small scales, since annotating temporal boundaries for all instances in videos is labor-intensive and time-consuming. Our UnAV-100 is the only dataset that combines both audio and visual signals to annotate instances, while others just utilize visual content in videos. Their audio tracks are usually very noisy and unrelated to the visual information, *e.g.*, background music and narrations, thus these datasets are not suitable for joint audio-visual video understanding. Besides, these benchmarks all focus on specific domains, such as human activities, sports, cooking, *etc*. By contrast, our UnAV-100 covers many different domains including human/music/sport/animal/nature, *etc*., which helps machines to understand more diverse audio-visual scenes in the wild.

2. Implementation Details

Feature Extraction. The visual features are extracted using two-stream I3D [2], which inputs a set of 24 RGB and optical flow frames extracted at 25 fps. Each frame is first resized such that the shortest side is 256 pixels, and then the center region is cropped to 224 × 224. A 1024-d RGB or flow feature vector is obtained from the final convolutional layer of the corresponding branch of I3D. Then, the two vectors are concatenated producing 2048-d features for each stack of 24 frames. The audio features are extracted using VGGish [5]. The input is a 96 × 64 log mel-scaled spectrogram extracted for each 0.96s segment, which is obtained by applying *Short-Time Fourier Transform* on a 16 kHz mono audio track. Then, a 128-d feature vector can be obtained after an activation function and before a classification layer. Here, we use 24 frames for each visual segment to temporally match with the input of the audio modality as $\frac{24}{25} = 0.96$.

Network Architecture. In the cross-modal pyramid transformer encoder, the number of attention heads is 4 in both uni-modal and cross-modal blocks. The temporal downsampling operation is realized by using a single depth-wise 1D convolution as in [10]. For temporal dependency modeling, the output dimension is converted as the shape of input to formulate it as a plug-and-play operation, and we just apply this operation once in our model.

Reproducibility. All our models are trained on a single 32GB NVIDIA Tesla V100 GPU and implemented in PyTorch deep-learning framework. During inference, we evaluate the performances of our method on the test set of our UnAV-100 and use the best models on the validation set.

3. Ablation Study

Position Encoding. We explore the impact of position encoding in our transformer encoder. As shown in Tab. 2, adding position embeddings can improve the performance by 0.8% in average mAP, even though the temporal convolutions (*i.e*., the projection layer and downsampling operations) already leak the location information as pointed out in [9, 10].

Loss Weight. We also provide the ablation study on the loss weight λ in our loss function. We train the model using different loss weights $\lambda \in [0.2, 0.5, 1, 2, 5]$, and report the results in Tab. 3. It can be seen that the default value $\lambda = 1$ can yield the best performance.

Feature Stride. In our experiments, we use stride=8 with a sliding window of 24 frames by default when extracting visual and audio features. Here, we study the performance variation using different feature strides in Tab. 4. Reducing the temporal feature resolution (*i.e*., larger strides, 16/24) leads to obvious performance degradation, which is intuitively reasonable since the model might fail to detect many short audio-visual events at a low temporal resolution.

PE	0.5	0.6	0.7	0.8	0.9	Avg.
✓	**50.6**	44.8	**39.8**	32.4	21.1	**47.8**
	49.5	**45.1**	39.7	**32.8**	**21.9**	47.0

Table 2. Ablation study on position encoding (PE).

λ	0.5	0.6	0.7	0.8	0.9	Avg.
0.2	49.9	45.0	39.6	32.2	20.7	46.9
0.5	50.1	45.4	39.8	32.3	21.2	47.3
1	**50.6**	**45.8**	39.8	32.4	21.1	**47.8**
2	49.8	45.3	**40.2**	**33.0**	**22.4**	47.2
5	49.0	44.7	39.2	32.3	22.2	46.4

Table 3. Ablation study on loss weight λ.

Stride	0.5	0.6	0.7	0.8	0.9	Avg.
8	**50.6**	**44.8**	**39.8**	32.4	21.1	**47.8**
16	48.9	44.6	39.0	**32.9**	**21.8**	46.7
24	49.7	44.7	38.5	31.0	20.9	47.0

Table 4. Ablation study on temporal feature stride.

T_{max}	0.5	0.6	0.7	0.8	0.9	Avg.
192	49.9	45.2	39.7	32.6	21.7	47.0
224	**50.6**	**45.8**	39.8	32.4	21.1	**47.8**
256	49.6	45.3	**39.9**	**33.1**	**22.3**	47.2

Table 5. Ablation study on maximum input sequence length.

SD	CD	0.5	0.6	0.7	0.8	0.9	Avg.
		48.5	43.4	36.9	29.9	20.2	45.8
✓		49.5	**45.3**	39.7	32.6	21.2	46.9
	✓	49.5	44.8	39.7	**32.7**	**21.7**	46.8
✓	✓	**50.6**	44.8	**39.8**	32.4	21.1	**47.8**

Table 6. Ablation study on dependency modeling. SD: simultaneous dependency branch; CD: consecutive dependency branch.

Maximum Input Sequence Length. Furthermore, we vary the length of the maximum input sequences of our model, and the results are provided in Tab. 5. We can observe that our model has quite stable results when using different T_{max}, and $T_{max} = 224$ gets the best results.

Dependency Modeling. Since the two branches of temporal dependency modeling aim to capture different correlations between events within a video, we run an ablation by removing each of the branches and show the results in Tab. 6. It indicates that applying each branch separately also leads to improvement, and the best result can be achieved by combing both branches to model simultaneous and consecutive dependencies at the same time.

4. Experiments on Existing TAL Dataset

We also conduct experiments on THUMOS14 dataset [6], a widely-used dataset for temporal action localization. The evaluation results on THUMOS14 test set using only visual input are provided in Tab. 7. We use the same strategy to extract features on THUMOS14 as used on UnAV-100 for both methods to keep a fair comparison. We can see that our model outperforms ActionFormer [10] by a large margin (+3.1% mAP at tIoU=0.5), even without the cross-modal fusion strategy. Besides, we tried to only use the audio modality in THUMOS14 to locate actions, but got very bad results (just 4.3% average mAP) on both models, which indicates that the audio tracks in THUMOS14 are quite noisy and cannot provide useful information.

Method	0.3	0.4	0.5	0.6	0.7	Avg.
ActionFormer [10]	73.4	67.5	57.6	47.6	33.7	56.0
Ours	**74.8**	**70.1**	**60.7**	**48.1**	**34.0**	**57.5**

Table 7. Experiments on THUMOS14 dataset with only visual modality as input (mAP@[0.3:0.1:0.7] is reported).

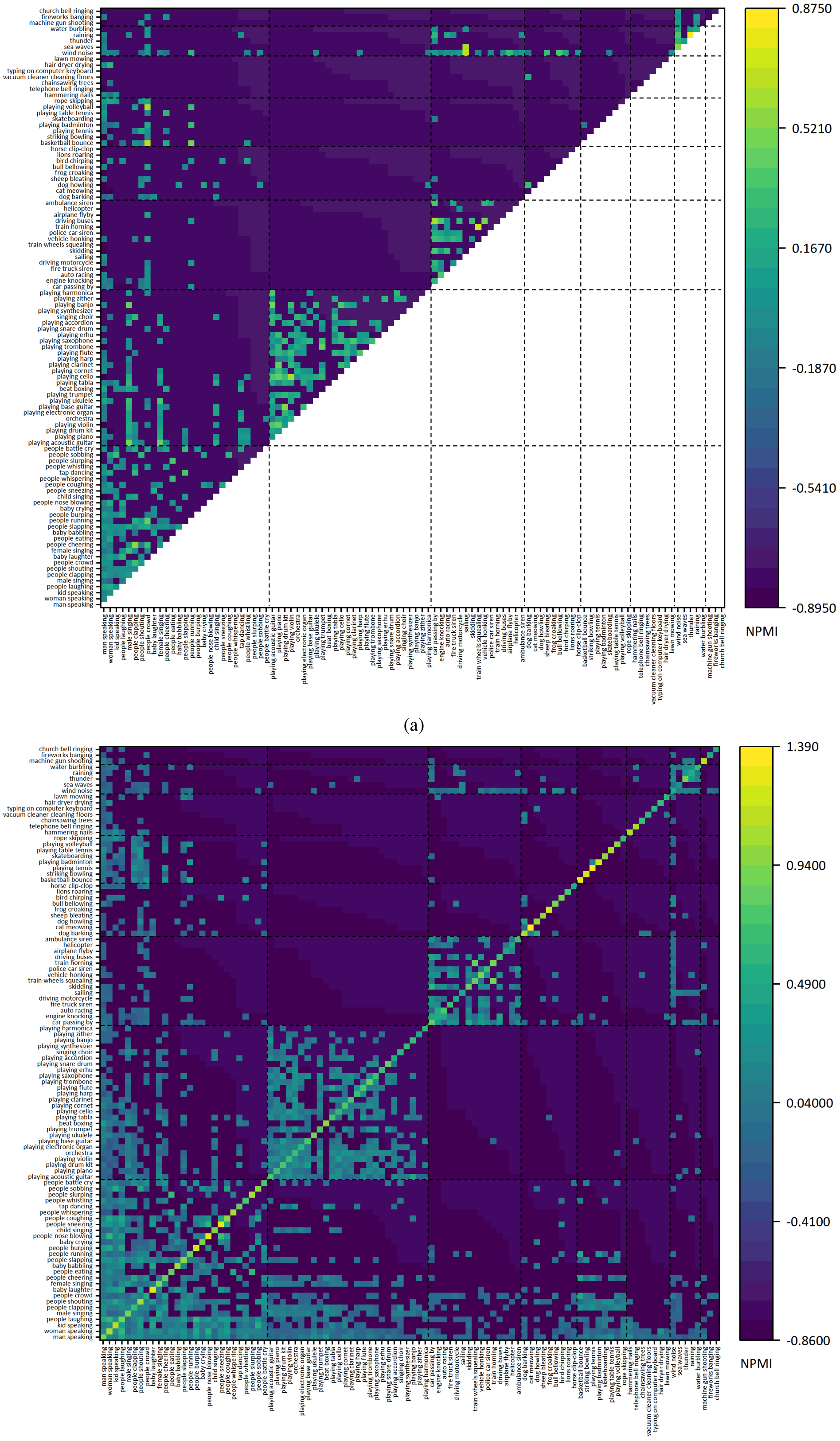
5. More Qualitative Results

More qualitative results are presented in Fig 3, which includes the prediction results of our model variants using different modalities as input. Generally speaking, cross-modal perception encourages the model to obtain more correct localization results. For example, Fig. 3(a) refers to the relatively constant visual information versus dramatically changing audio signals. By integrating both modalities, the model can better judge the event boundaries. Besides, our audio-visual model can also get promising performance in some complex audio-visual scenarios, as in Fig. 3(c) and Fig. 3(d), where many audio-visual events occur concurrently or over very short periods of time.

6. Discussion

Limitations. There is still a wide scope for exploration and improvement on the basis of our work. For instance, our dataset is limited to a temporal localization task. We will explore other audio-visual problems, such as representation learning and sound source localization in real-life and complex scenarios in our subsequent study. Besides, although our model can obtain a promising performance, as a baseline, its capability is still limited in some complex situations. For example, in Fig. 3(c), the model gets an incorrect boundary of the *dog barking* event when the barking brown dog is out of the screen and a non-barking black one can be seen. This indicates that our model might fail to effectively filter out interference information for such a difficult case. And the model might also fail to predict precise boundaries when one modality persists while another disappears for a short period of time (*e.g.*, the event of *vacuum cleaner cleaning floors* in Fig. 3(c)). In addition, for some instant events with very short duration (*e.g.*, *basketball bounce* in Fig. 3(d)), our model might get unsatisfactory results. Overall, dense-localizing audio-visual events is inherently a very challenging task, and it requires the model to have a strong fine-grained cross-modal understanding ability. Therefore, more advanced models that could better solve the above difficulties are expected to boost performance further. We hope our work as the first attempt at untrimmed audio-visual video understanding can inspire more people to explore the field.

Ethic concerns and biases. Our UnAV-100 is sourced from VGGSound dataset [3] that has already tried to mitigate ethical issues. During data collection, we made further efforts to manually check all videos to avoid mature, sensitive, or offensive content. Besides, our UnAV-100 follows the natural distribution of instances present on the website, which may reflect some biases in topics. For example, there are more *man/woman speaking* events than other categories. Efforts have been made to mitigate such imbalance.



(a)

(b)

Figure 2. NPMI of the pairs of simultaneous (a) and consecutive (b) audio-visual events in our UnAV-100 dataset. In (b), the horizontal axis shows the first event, and the vertical axis shows the second subsequent event. The event categories are grouped by domains.

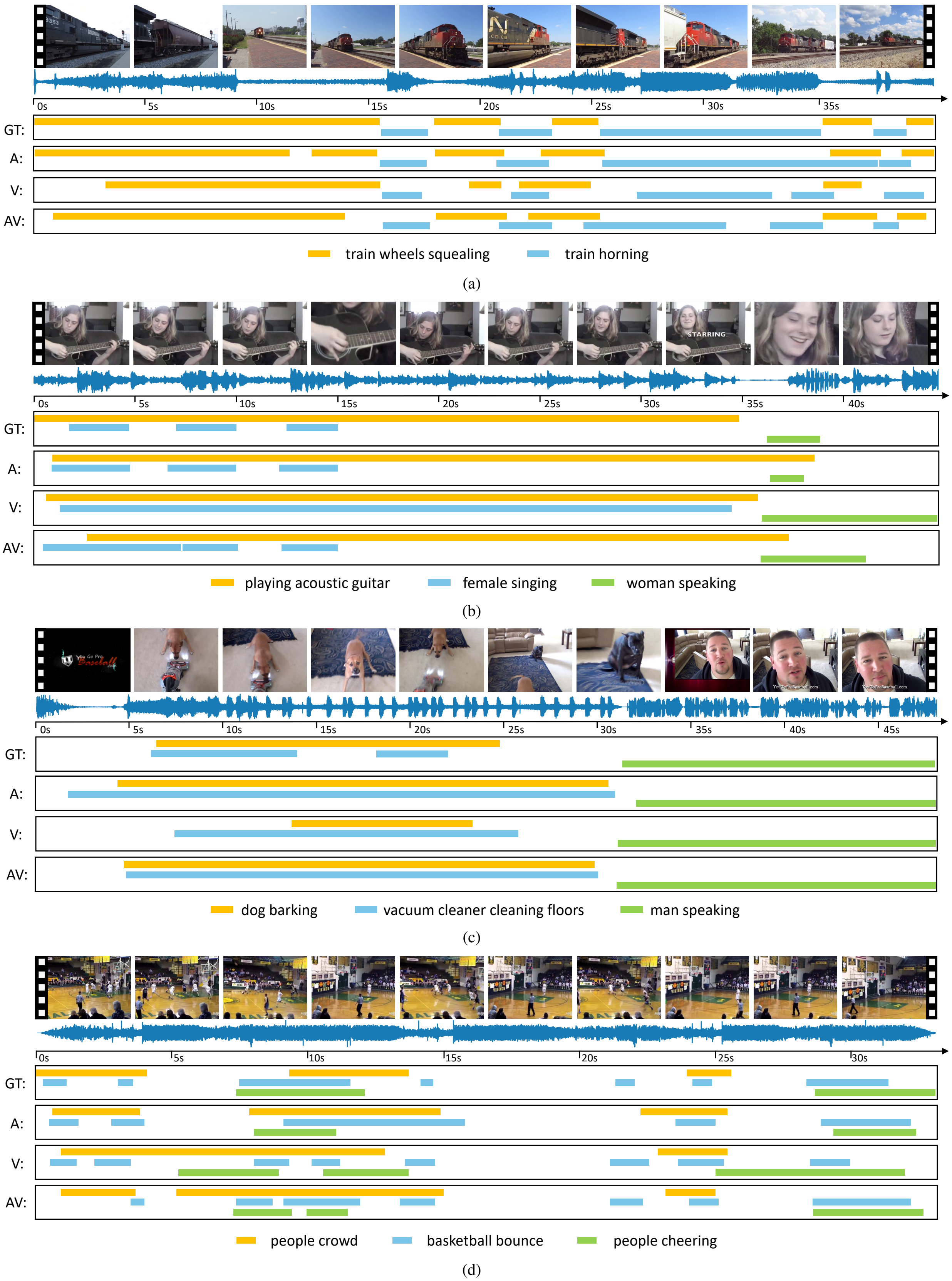


Figure 3. More qualitative results on the UnAV-100 test set. GT: ground truth, A: the prediction of the audio-only variant, V: the prediction of the visual-only variant, AV: the prediction of our audio-visual model. We show boundaries with the highest overlap with ground truth.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 3

[4] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990. 1

[5] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 2

[6] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23, 2017. 1, 3

[7] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 1

[8] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 1

[9] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[10] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 2, 3