Weakly-Supervised Audio-Visual Video Parsing Toward Unified Multisensory Perception

Yapeng Tian[1], Dingzeyu Li[2], and Chenliang Xu[1]
[1]University of Rochester [2]Adobe Research

1. Introduction

Utilizing and learning from both auditory and visual modalities is an emerging research topic. Recent years have seen progress in learning representations [1, 2, 6], separating visually indicated sounds [3, 14], spatially localizing visible sound sources [8, 12], and temporally localizing audio-visual synchronized segments [12]. However, past approaches usually assume audio and visual data are always correlated or even temporally aligned. In practice, when we analyze the video scene, many videos have audible sounds, which originate outside of the FoV, leaving no visual correspondences, but still contribute to the overall understanding, such as out-of-screen running cars and a narrating person. Such examples are so ubiquitous, which leads us to some basic questions: what video events are audible, visible, and "audi-visible," where and when are these events inside of a video, and how can we effectively detect them?

To answer the above questions, we pose and try to tackle a fundamental problem: *audio-visual video parsing* that recognizes event categories bind to sensory modalities, and meanwhile, finds temporal boundaries of when such an event starts and ends (see Fig. 1). However, learning a fully supervised audio-visual video parsing model requires densely annotated event modality and category labels with corresponding event onsets and offsets, which will make the labeling process extremely expensive and time-consuming. To avoid tedious labeling, we explore weakly-supervised learning for the task, which only requires sparse labeling on the presence or absence of video events. The weak labels are easier to annotate and can be gathered in a large scale from web videos.

We formulate the weakly-supervised audio-visual video parsing as a Multimodal Multiple Instance Learning (MMIL) problem and propose a new framework to solve it. Concretely, we use a new hybrid attention network (HAN) for leveraging unimodal and cross-modal temporal contexts simultaneously. We develop an attentive MMIL pooling method for adaptively aggregating useful audio and visual content from different temporal extent and modalities. Furthermore, we discover modality bias and noisy label issues and alleviate them with an individual-guided learning mechanism and label smoothing [7], respectively.

To facilitate our investigations, we collect a *Look, listen, and Parse* (LLP) dataset that has $11,849$ YouTube video clips from 25 event categories. We label them with sparse video-level event labels for training. For evaluation, we label a set of precise labels, including event modalities, event categories, and their temporal boundaries. Experimental results show that it is tractable to learn audio-visual video parsing even with video-level weak labels. Our proposed HAN model can effectively leverage multimodal temporal contexts. Furthermore, modality bias and noisy label problems can be addressed with the proposed individual learning strategy and label smoothing, respectively.

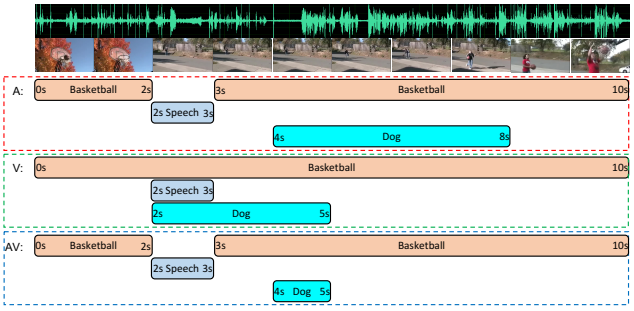

Figure 1: Our audio-visual video parsing model aims to parse a video into different audio (audible), visual (visible), and audio-visual (audi-visible) events with correct categories and boundaries. A dog in the video visually appears from 2nd second to 5th second and make barking sounds from 4th second to 8th second. So, we have audio event (4s-8s), visual event (2s-5s), and audio-visual event (4s-5s) for the *Dog* event category.

2. Dataset and Problem

LLP: The Look, Listen and Parse Dataset To the best of our knowledge, there is no existing dataset that is suitable for us. Thus, we introduce a *Look, Listen, and Parse* dataset for audio-visual video scene parsing, which contains 11,849 YouTube video clips spanning over 25 categories for a total of 32.9 hours collected from the AudioSet. A wide range of video events (*e.g.*, human speaking, singing, baby crying, dog barking, violin playing, and car running, and vacuum cleaning *etc.*) from diverse domains (*e.g.*, human activities, animal activities, music performances, vehicle sounds, and domestic environments) are included in the dataset. Each video is $10s$ long and has at least $1s$ audio or visual events. There are 7,202 videos that contain events from more than one event categories and per video has averaged 1.64 different event categories. To evaluate audio-visual scene parsing performance, we annotate individual audio and visual events with second-wise temporal boundaries for randomly selected 1,849 videos from the LLP dataset. Note that the audio-visual event labels can be derived from the audio and visual

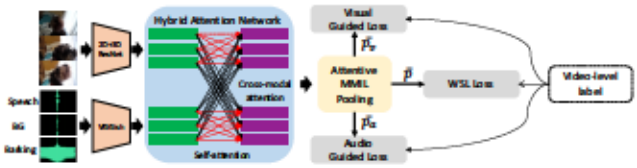

Figure 2: Our audio-visual video parsing framework. It uses pre-trained CNNs to extract snippet-level audio and visual features and leverages multimodal temporal contexts with the proposed hybrid attention network (HAN). For each snippet, we will predict both audio and visual event labels from the aggregated features by the HAN. Attentive MMIL pooling is utilized to adaptively predict video-level event labels for weakly-supervised learning (WSL) and individual guided learning is devised to mitigate the modality bias issue.

event labels. Finally, we have totally 6,626 event annotations, including 4,131 audio events and 2,495 visual events for the 1,849 videos. Merging the individual audio and visual labels, we obtain 2,488 audio-visual event annotations.

Audio-Visual Video Parsing with Weak Labels We define the *Audio-Visual Video Parsing* as a task to group video segments and parse a video into different temporal audio, visual, and audio-visual events associated with semantic labels. Since event boundary in the LLP dataset was annotated at second-level, video events will be parsed at scene-level not object/instance level in our experimental setting. Concretely, given a video sequence containing both audio and visual tracks, we divide it into T non-overlapping audio and visual snippet pairs $\{V_t, A_t\}_{t=1}^{T}$, where each snippet is $1s$ long and V_t and A_t denote visual and audio content in the same video snippet, respectively. Let $\mathbf{y}_t = \{(y_t^a, y_t^v, y_t^{av}) | [y_a^t]_c, [y_v^t]_c, [y_{av}^t]_c \in \{0,1\}, c = 1, ..., C\}$ be the event label set for the video snippet $\{V_t, A_t\}$, where c refers to the c-th event category and y_t^a, y_t^v, and y_t^{av} denote audio, visual, and audio-visual event labels, respectively. Here, we have a relation: $y_t^{av} = y_t^a * y_t^v$, which means that audio-visual events occur only when there exists both audio and visual events at the same time and from the same event categories. In this work, we explore the audio-visual video parsing in a weakly-supervised manner. We only have video-level labels for training, but will predict precise event label sets for all video snippets during testing, which makes the weakly-supervised audio-visual video parsing be a multimodal multiple instance learning (MMIL) problem.

3. Method

Audio-Visual Video Parsing Framework Our framework, as illustrated in Fig. 2, has three main modules: audio and visual feature extraction, multimodal temporal modeling, and attentive MMIL pooling. Given a video sequence with T audio and visual snippet pairs $\{V_t, A_t\}_{t=1}^{T}$, we first use pre-trained visual and audio models to extract snippet-level visual features: $\{f_v^t\}_{t=1}^{T}$ and audio features: $\{f_a^t\}_{t=1}^{T}$, respectively. Taking extracted audio and visual features as inputs, we use two hybrid attention networks as the multimodal temporal modeling module to leverage unimodal and cross-modal temporal contexts and obtain updated visual features $\{\hat{f}_v^t\}_{t=1}^{T}$ and audio features $\{\hat{f}_a^t\}_{t=1}^{T}$. To predict audio and visual instance-level labels and make use of the video-level weak labels, we address the MMIL problem with a novel attentive MMIL pooling module outputting video-level labels.

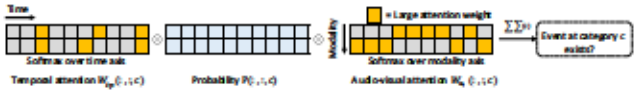

Figure 3: Attentive MMIL Pooling. For event category c, temporal and audio-visual attention mechanisms will adaptively select informative event predictions crossing temporal and modality axes, respectively, for predicting whether there is an event at the category.

Hybrid Attention Network At each time step t, a hybrid attention function g in HAN will be learned from audio and visual features: $\{f_a^t, f_v^t\}_{t=1}^{T}$ to update f_a^t and f_v^t, respectively. The updated audio feature $\hat{f}_a^t$ and visual feature $\hat{f}_v^t$ can be computed as: $\hat{f}_a^t = g(f_a^t, \boldsymbol{f}_a, \boldsymbol{f}_v) = f_a^t + g_{sa}(f_a^t, \boldsymbol{f}_a) + g_{ca}(f_a^t, \boldsymbol{f}_v)$ and $\hat{f}_v^t = g(f_v^t, \boldsymbol{f}_a, \boldsymbol{f}_v) = f_v^t + g_{sa}(f_v^t, \boldsymbol{f}_v) + g_{ca}(f_v^t, \boldsymbol{f}_a)$, where $\boldsymbol{f}_a = [f_a^1; ...; f_a^T]$ and $\boldsymbol{f}_v = [f_v^1; ...; f_v^T]$; g_{sa} and g_{ca} are self-attention and cross-modal attention functions, respectively; skip-connections can help preserve the identity information from the input sequences. The two attention functions are formulated with the same computation mechanism. With $g_{sa}(f_a^t, \boldsymbol{f}_a)$ and $g_{ca}(f_a^t, \boldsymbol{f}_v)$ as examples, they are defined as: $g_{sa}(f_a^t, \boldsymbol{f}_a) = \sum_{t=1}^{T} w_t^{sa} f_a^t = softmax(\frac{f_a^t \boldsymbol{f}_a'}{\sqrt{d}}) \boldsymbol{f}_a$ and $g_{ca}(f_a^t, \boldsymbol{f}_v) = \sum_{t=1}^{T} w_t^{ca} f_v^t = softmax(\frac{f_a^t \boldsymbol{f}_v'}{\sqrt{d}}) \boldsymbol{f}_v$, where the scaling factor d is equal to the audio/visual feature dimension and $(\cdot)'$ denotes the transpose operator. Clearly, the self-attention and cross-modal attention functions in HAN will assign large weights to snippets, which are similar to the query snippet containing the same video events within the same modality and cross different modalities.

Attentive MMIL Pooling To achieve audio-visual video parsing, we predict all event labels for audio and visual snippets from temporal aggregated features: $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^{T}$. We use a shared fully-connected layer to project audio and visual features to different event label space and adopt a sigmoid function to output probability for each event category: $p_a^t = sigmoid(FC(\hat{f}_a^t))$ and $p_v^t = sigmoid(FC(\hat{f}_v^t))$, where p_a^t and p_v^t are predicted audio and visual event probabilities at timestep t, respectively.

Since audio-visual events only occur when sound sources are visible and their sounds are audible, the audio-visual event probability p_{av}^t can be derived from individual audio and visual predictions: $p_{av}^t = p_a^t * p_v^t$. If we have direct

supervisions for all audio and visual snippets from different time steps, we can simply learn the audio-visual video parsing network in a fully-supervised manner. However, in this MMIL problem, we can only access a video-level weak label $\bar{\boldsymbol{y}}$ for all audio and visual snippets: $\{A_t, V_t\}_{t=1}^T$ from a video. To learn our network with weak labels, as illustrated in Fig. 3, we propose a attentive MMIL pooling method to predict video-level event probability: $\bar{\boldsymbol{p}}$ from $\{p_a^t, p_v^t\}_{t=1}^T$. Concretely, the $\bar{\boldsymbol{p}}$ is computed by: $\bar{\boldsymbol{p}} = \sum_{t=1}^T \sum_{m=1}^M (W_{tp} \odot W_{av} \odot P)[t, m, :]$, where $\odot$ denotes element-wise multiplication; m is a modality index and $M = 2$ refers to audio and visual modalities; W_{tp} and W_{av} are temporal attention and audio-visual attention tensors predicted from $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^T$, respectively, and P is the probability tensor built by $\{p_a^t, p_v^t\}_{t=1}^T$ and we have $P(t, 0, :) = p_a^t$ and $P(t, 1, :) = p_v^t$. To compute the two attention tensors, we first compose an input feature tensor F, where $F(t, 0, :) = \hat{f}_a^t$ and $F(t, 1, :) = \hat{f}_v^t$. Then, two different FC layers are used to transform the F into two tensors: F_{tp} and F_{av}, which has the same size as P. To adpatively select most informative snippets for predicting probabilities of different event categories, we assign different weights to snippets at different time steps with a temporal attention mechanism: $W_{tp}[:, m, c] = softmax(F_{tp}[:, m, c])$, where $m = 1, 2$ and $c = 1, \ldots, C$. Accordingly, we can adaptively select most informative modalities with the audio-visual attention tensor: $W_{av}[t, :, c] = softmax(F_{av}[t, :, c])$, where $t = 1, \ldots, T$ and $c = 1, \ldots, C$. The snippets within a video from different temporal steps and different modalities may have different video events. The proposed attentive MMIL pooling can well model this observation with the tensorized temporal and multimodal attention mechanisms.

With the predicted video-level event probability $\bar{\boldsymbol{p}}$ and the ground truth label $\bar{\boldsymbol{y}}$, we can optimize the proposed weakly-supervised learning model with a binary cross-entropy loss function: $\mathcal{L}_{wsl} = CE(\bar{\boldsymbol{p}}, \bar{\boldsymbol{y}}) = -\sum_{c=1}^C \bar{\boldsymbol{y}}[c] log(\bar{\boldsymbol{p}}[c])$.

Alleviating Modality Bias and Label Noise However, it usually enforces models to only identify discriminative patterns in the training data, which was observed in previous weakly-supervised MIL problems [9, 10]. In our MMIL problem, the issue becomes even more complicated since there are multiple modalities and different modalities might not contain equally discriminative information. With weakly-supervised learning, the model tends to only use information from the most discriminative modality but ignore another modality, which can probably achieve good video classification performance but terrible video parsing performance on the events from ignored modality and audio-visual events. Since a video-level label contains all event categories from audio and visual content within the video, to alleviate the modality bias in the MMIL, we propose to use explicit supervisions to both modalities with a guided loss: $\mathcal{L}_g = CE(\bar{\boldsymbol{p}}_a, \bar{\boldsymbol{y}}_a) + CE(\bar{\boldsymbol{p}}_v, \bar{\boldsymbol{y}}_v)$, where $\bar{\boldsymbol{y}}_a = \bar{\boldsymbol{y}}_v = \bar{\boldsymbol{y}}$, and $\bar{\boldsymbol{p}}_a = \sum_{t=1}^T (W_{tp} \odot P)[t, 0, :]$ and $\bar{\boldsymbol{p}}_v = \sum_{t=1}^T (W_{tp} \odot P)[t, 1, :]$ are video-level audio and visual event probabilities, respectively.

However, not all video events are audio-visual events, which means that an event occurred in one modality might not occur in another modality and then the corresponding event label will be label noise for one of the two modalities. Thus, the guided loss: $\mathcal{L}_g$ suffers from noisy label issue. For the example shown in Fig. 2, the video-level label is {*Speech*, *Dog*} and the video-level visual event label is only {*Dog*}. The {*Speech*} will be a noisy label for the visual guided loss. To handle the problem, we lower the confidence of positive labels with smoothing $\bar{\boldsymbol{y}}$ and generate smoothed labels: $\bar{\boldsymbol{y}}_a$ and $\bar{\boldsymbol{y}}_v$. They are formulated as: $\bar{\boldsymbol{y}}_a = (1 - \epsilon_a)\bar{\boldsymbol{y}} + \frac{\epsilon_a}{K}$ and $\bar{\boldsymbol{y}}_v = (1 - \epsilon_v)\bar{\boldsymbol{y}} + \frac{\epsilon_v}{K}$, where $\epsilon_a, \epsilon_v \in [0, 1)$ are two confidence parameters to balance the event probability distribution and a uniform distribution: $u = \frac{1}{K}$ $(K > 1)$. For a noisy label at event category c, when $\bar{\boldsymbol{y}}[c] = 1$ and real $\bar{\boldsymbol{y}}_a[c] = 0$, we have $\bar{\boldsymbol{y}}[c] = (1 - \epsilon_a)\bar{\boldsymbol{y}}[c] + \epsilon_a > (1 - \epsilon_a)\bar{\boldsymbol{y}} + \frac{\epsilon_a}{K} = \bar{\boldsymbol{y}}_a[c]$ and the smoothed label will become more reliable. The proposed method is inspired by the label smoothing [11] technique. Different from the past methods, we use smoothed labels to mitigate label noise occurred in the individual guided learning. Our final model is optimized with the two loss terms: $\mathcal{L} = \mathcal{L}_{wsl} + \mathcal{L}_g$.

4. Experiments

Baselines. Since there are no existing methods to address the audio-visual video parsing, we design several baselines based on previous state-of-the-art weakly-supervised sound detection [13], temporal action localization [4], and audio-visual event localization [12] methods to validate our framework. For fair comparisons, the compared approaches use the same audio and visual features as our method.

Evaluation Metrics. To comprehensively measure the performance of different methods, we evaluate them on parsing all types of events (individual audio, visual, and audio-visual events) under both segment-level and event-level metrics. To evaluate overall audio-visual scene parsing performance, we also compute aggregated results, where Type@AV computes averaged audio, visual, and audio-visual event evaluation results and Event@AV computes the F-score considering all audio and visual events for each sample rather than directly averaging results from different event types as the Type@AV. We use both segment-level and event-level F-scores [5] as metrics. The segment-level metric can evaluate snippet-wise event labeling performance. For computing event-level F-score results, we extract events with concatenating positive consecutive snippets in the same event categories and compute the event-level F-score based on mIoU = 0.5 as the threshold.

Quantitative Comparison The quantitative results are shown in Tab. 1. We can see that our method outperforms

Table 1: Audio-visual video parsing accuracy (%) of different methods on the LLP test dataset. These methods all use the same audio and visual features as inputs for a fair comparison. The top-1 results in each line are highlighted.

Event type	Methods	Segment-level	Event-level
Audio	TALNet [13]	50.0	41.7
	AVE [12]	47.2	40.4
	Ours	**56.5**	**45.0**
Visual	CMCS [4]	48.1	45.1
	AVE [12]	37.1	34.7
	Ours	**53.6**	**47.5**
Audio-Visual	AVE [12]	35.4	31.6
	Ours	**47.7**	**38.6**
Type@AV	AVE [12]	39.9	35.5
	Ours	**52.6**	**43.7**
Event@AV	AVE [12]	41.6	36.5
	Ours	**53.5**	**43.3**

Table 2: Ablation study on learning mechanism, attentive MMIL pooling, hybrid attention network, and handling noisy labels. Segment-level parsing results are shown.

Loss	MMIL Pooling	Temporal Net	Handle Noisy Label	Audio	Visual	Audio-Visual	Type@AV	Event@AV
$\mathcal{L}_{wsl}$	Attentive	×	×	**56.9**	16.4	17.2	30.2	43.3
$\mathcal{L}_g$	Attentive	×	×	42.3	43.9	34.5	40.3	42.0
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	×	×	45.1	**51.7**	**35.0**	**44.0**	**48.9**
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Max	×	×	31.6	43.6	22.5	32.6	39.1
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Mean	×	×	40.2	43.2	**35.0**	39.5	39.7
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	×	×	**45.1**	**51.7**	**35.0**	**44.0**	**48.9**
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	×	×	45.1	51.7	35.0	44.0	48.9
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	GRU	×	52.0	49.4	39.0	46.8	51.0
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	Transformer	×	51.2	52.6	40.7	48.2	51.5
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	HAN	×	**53.9**	**52.8**	**43.5**	**50.1**	**53.0**
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	HAN	×	53.9	52.8	43.5	50.1	53.0
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	HAN	Bootstrap [7]	45.0	50.9	35.0	43.6	48.2
$\mathcal{L}_{wsl} + \mathcal{L}_g$	Attentive	HAN	ours	**56.5**	**53.6**	**47.7**	**52.6**	**53.5**

compared approaches on all audio-visual video parsing subtasks under both the segment-level and event-level metrics, which demonstrates that our network can predict more accurate snippet-wise event categories with more precise event onsets and offsets for testing videos on the LLP dataset.

Individual Guided Learning. From Tab. 2, we observe that the model without individual guided learning can achieve pretty good performance on audio event parsing but incredibly bad visual parsing results leading to terrible audio-visual event parsing; w/ only $\mathcal{L}_g$ model can achieve both reasonable audio and visual event parsing results; our model trained with both $\mathcal{L}_{wsl}$ and $\mathcal{L}_g$ outperforms model train without and with only $\mathcal{L}_g$. The results indicate that the model trained only $\mathcal{L}_{wsl}$ find discriminative information from mostly sounds and visual information is not well-explored and the individual learning can effectively handle the modality bias issue.

Attentive MMIL Pooling. Our Attentive MMIL Pooling (see Tab. 2) is superior over the both compared Max pooling and Mean pooling methods. Our attentive MMIL pooling allows assigning different weights to audio and visual snippets within a video bag for each event category, thus can adaptively discover useful snippets and modalities.

Hybrid Attention Network. We compare our HAN with two temporal networks: GRU and Transformer and a model without temporal modeling in Tab. 2. Clearly, our HAN is more effective in leveraging multimodal temporal contexts.

Noisy Label. Tab. 2 also shows results of our model without handling the noisy label, with Bootstrap [7] and our label smoothing-based method. Our method with reducing confidence for potential false positive labels can help to learn a more robust model with improved video parsing results.

Acknowledgments

This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *ICCV*, 2017. 1

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 1

[3] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, pages 35–53, 2018. 1

[4] D. Liu, T. Jiang, and Y. Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, 2019. 3, 4

[5] A. Mesaros, T. Heittola, and T. Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016. 3

[6] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1

[7] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 1, 4

[8] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1

[9] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 3

[10] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, 2014. 3

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[12] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 1, 3, 4

[13] Y. Wang, J. Li, and F. Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP*, pages 31–35. IEEE, 2019. 3, 4

[14] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *ECCV*, 2018. 1