Visual Scene Graphs for Audio Source Separation
Supplementary Materials

Moitreya Chatterjee[1] Jonathan Le Roux[2] Narendra Ahuja[1] Anoop Cherian[2]
[1]University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
[2]Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA
metro.smiles@gmail.com leroux@merl.com n-ahuja@illinois.edu cherian@merl.com

1. Summary of Supplementary Results

In this supplementary document, we first provide additional details about the challenging *Audio Separation in the Wild (ASIW)* dataset, which we introduce in this work. These details are provided in Section 2. This is followed by the details of our neural network architecture in Section 3. Finally, we present qualitative results of our *visually-guided* audio source separation experiments in Section 4, including a user assessment study.

To summarize, below is the list of additional results that we provide:

1. ASIW dataset details.
2. Network architecture details.
3. Additional qualitative results.

2. ASIW Dataset Details

Most prior approaches in visually-guided sound source separation report performances solely in the setting of separating the sounds of musical instruments [3, 15, 14, 2]. However, musical instruments often have very characteristic sounds and thereby the range of variability within a particular instrument category is limited. Moreover, the videos featured in these datasets are often recorded professionally in rather controlled environments, such as an auditorium. Such videos however, may not capture the variety of sounds that we come across in daily-life settings. In order to fill this void, this work introduces the *Audio Separation in the Wild (ASIW)* dataset.

ASIW is adapted from the recently introduced large-scale AudioCaps dataset [6], which contains 49,838 training, 495 validation, and 975 test videos crawled from the AudioSet dataset [4], each of which is about 10s long. These videos have been carefully captioned manually (by English-speaking Amazon Mechanical Turkers – AMTs). In comparison to other video captioning datasets (such as MSVD or MSRVTT), AudioCaps captions are particularly focused on describing auditory events in the video; which motivated us to consider this dataset for the task of visually-guided sound source separation.

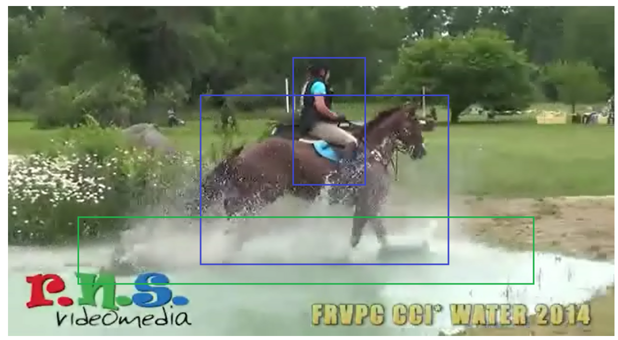

Figure 1. A sample frame from a video in the ASIW dataset, showing the principal object (water, highlighted by a green box) and a set of interacting objects (*horse* with a *rider*, highlighted by a blue box).

To adapt AudioCaps for our task, we manually construct a dictionary of 306 frequently occurring *auditory words* from the captions, such as: *splashing*, *flushing*, *eruptions*, or *giggling*. Another factor we considered in order to select this dictionary is the grounding that the words have in the video; which we call the *pricipal objects* in the main paper. The words in the dictionary are selected such that they have a corresponding principal object in the video generating the respective sound. The set of principal objects that we finally selected from AudioCaps consisted of 14 classes, namely: *baby*, *bell*, *birds*, *camera*, *clock*, *dogs*, *toilet*, *horse*, *man/woman*, *sheep/goat*, *telephone*, *trains*, *vehicle/car/truck*, *water*, and an additional *background* class, which encompasses words that usually do not consistently ground to a visible principal object in the video. For instance, *brushing* could ground to a person brushing his/her teeth with a toothbrush or could also map to a painter putting his/her strokes on a canvas. We construct the *principal object list* from the Visual Genome [8] classes. The number of videos in each of these classes is shown in Table 1. In Figure 1, we show a sample frame from a video in this dataset, highlighting the

Table 1. Number of videos for each of the principal object categories of ASIW dataset.

Baby	Bell	Birds	Camera	Clock	Dogs	Toilet	Horse	Man	Sheep	Telephone	Trains	Vehicle	Water
1616	151	2887	913	658	1407	838	385	6210	710	222	141	779	378

principal object (in green) - in this case *water* interacting with another object (in blue), viz. *a horse with a jockey*, to produce the auditory word *splashing*.

In Section 5, we list the full set of auditory words (in **bold-face** font), indicating alongside which principal object it is grounded to as well as its frequency in the captions associated with the dataset. While constructing the dataset, all principal object classes which consistently exhibit the same sound are treated as the same class and are indicated in the above list in the same row, separated by a forward-slash ('/'). For instance, although the class "clock" is different from the class "clock tower", visually, but since a possible sound emitted by both may be characterized as "donging", we treat them as equivalent principal objects. We intend to make this dataset publicly available for researchers in the community, upon the acceptance of this work.

3. Network Architecture Details

Our model, the *Audio Visual Scene Graph Segmenter* (AVSGS) has several components. Below, we list the key details of each of the components.

3.1. Feature Extractor

Our model commences with extracting features, corresponding to bounding boxes in the scene. In order to do so, we use a Faster R-CNN model [10], with a ResNet-101 [5] backbone pre-trained on the Visual Genome Dataset [8]. In order to obtain instrument features for the MUSIC dataset another detector [3] is trained on the the OpenImages dataset [7]. The former gives 2048-dimensional vectors, while the latter gives 512-dimensional vectors. In order to maintain consistency of feature dimensions across objects, we further encode the 2048-dimensional vectors into 512-dimensions through a 2-layer Multi-layer perceptron with Leaky ReLU activations (negative slope=0.2)

3.2. Graph Attention Network

Post the object detection and feature extraction, the scene-graph is constructed following the method laid out in the *Proposed Method* section of the paper. The scene graph is then processed by a *Graph Attention Network*, which has a cascade of the following three components:

Graph Attention Network Convolution: The Graph Attention Network Convolution (GATConv) [12] updates the node features of the graph based on the edge adjacency information by applying multi-head graph message-passing. We use 4 heads in the network and the dimension of the output feature of this network is 512.

Edge Convolution: Next, we employ Edge Convolutions [13] to capture pair-wise interactions, which take in a concatenated vector of 2 objects ($512 \times 2 = 1024$) and generates a 512-dimensional vector.

Pooling Layers: The final step of the *Graph Attention Network* consists of pooling these feature vectors [9] to obtain a single vector. We concatenate the embeddings obtained by Global Max and Average Pool to obtain this.

3.3. Recurrent Network

Our Recurrent Network is instantiated via a *Gated Recurrent Unit* (GRU) [1], whose input space and feature dimensions are 512-dimensional.

3.4. Audio Separator Network

A key component of our model is the audio separator network that takes as input a mixed audio track and produces a separated sound source as output, conditioned on a visual feature. The network roughly follows a U-Net [11] style architecture, with the visual feature being concatenated into the network at the bottleneck layer. The network has 7 convolution and 7 up-convolution layers, each with 4×4 filter dimensions and LeakyRELU activations with negative slope of 0.2. Additionally, there are skip connections between a pair of layers in the encoder and the decoder, with matching spatial resolution of their feature maps. The bottleneck layer has $2 \times 2 \times 512$ dimension and thus the visual feature vector obtained from the pre-processing above is tiled 2×2 times and then concatenated into the network at the bottleneck layer, along the channel dimension.

4. Qualitative Results

In this section, we present separated spectrogram visualizations obtained by our method versus competing baselines on both datasets, for a qualitative assessment by the reader. To this end, we show spectrogram separations for audio obtained from a mix of two different videos as well as separations on videos which have a mixture of multiple sound sources.

4.1. Qualitative Visualizations

From the qualitative visualizations presented in Figures 2, 3, 4, 5, 6, 11, 12, 13, 14, 15 we see that AVSGS is better able to separate the audio compared to competing baseline methods on ASIW and MUSIC respectively. We also notice that the separations obtained by AVSGS are more artifact free. Addtionally, in Figures 7, 8, 9, 10, 16, 17 we notice that AVSGS is adept at separating multiple sound

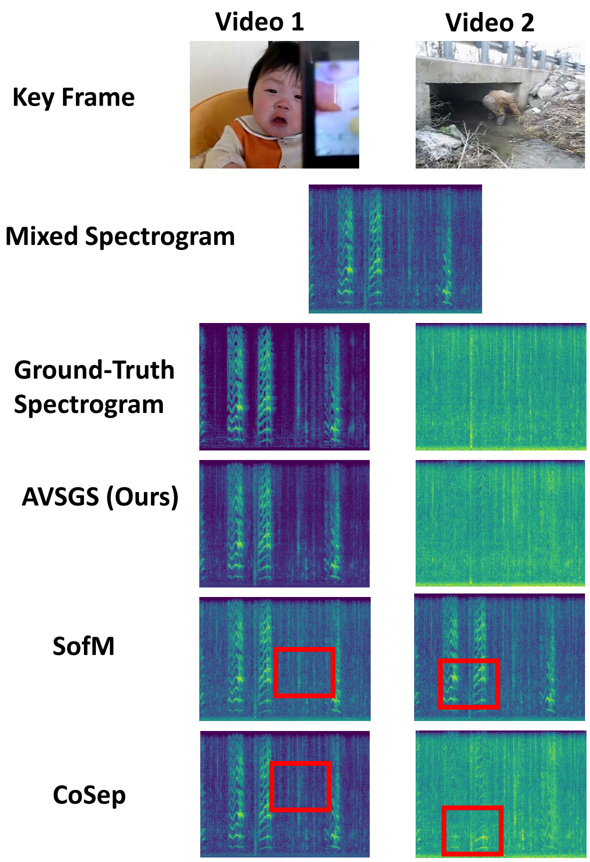

Figure 2. Qualitative separation results on a mixture of two ASIW videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

Table 2. Human preference score on samples from our method vs. Zhao *et al*. [14]

Datasets	Prefer ours
ASIW - Ours vs. [14]	**92%**
MUSIC - Ours vs. [14]	**83%**

sources from the same video, as reflected by the difference in the resultant separated spectrograms from the 2 sources.

4.2. Human Preference Evaluations

In order to subjectively assess the quality of audio source separation, we evaluated a randomly chosen subset of separated audio samples from AVSGS and our closest non-MUSIC-specific competitor SofM for human preferability, on both ASIW and MUSIC datasets. Table 2 reports these results and shows a clear preference of the evaluators, for our method over SofM on average 80-90% of the time.

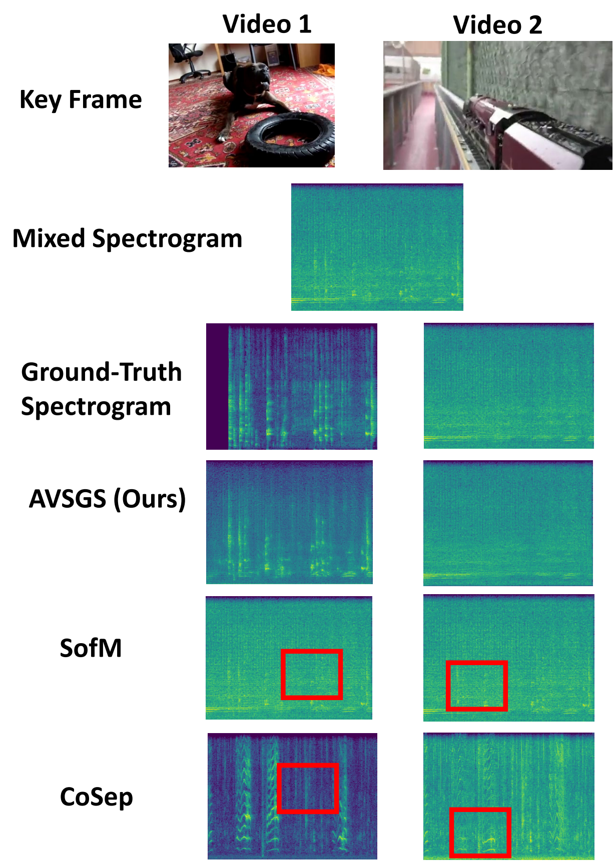

Figure 3. Qualitative separation results on a mixture of two ASIW videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

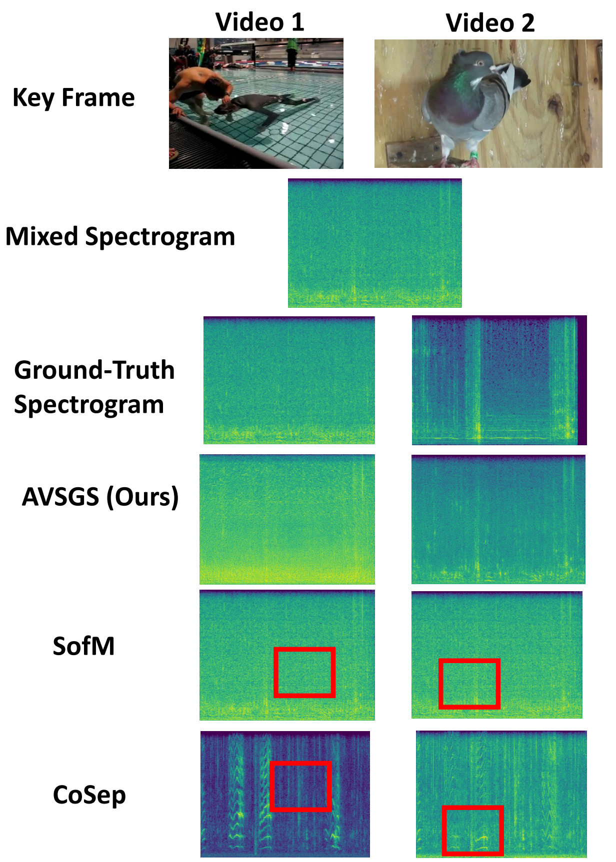

Figure 4. Qualitative separation results on a mixture of two ASIW videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

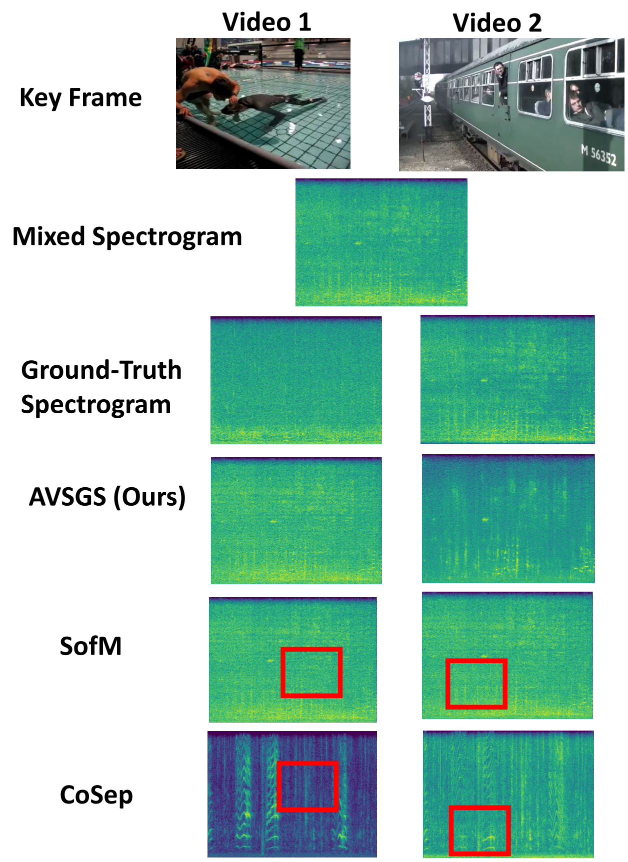

Figure 5. Qualitative separation results on a mixture of two ASIW videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

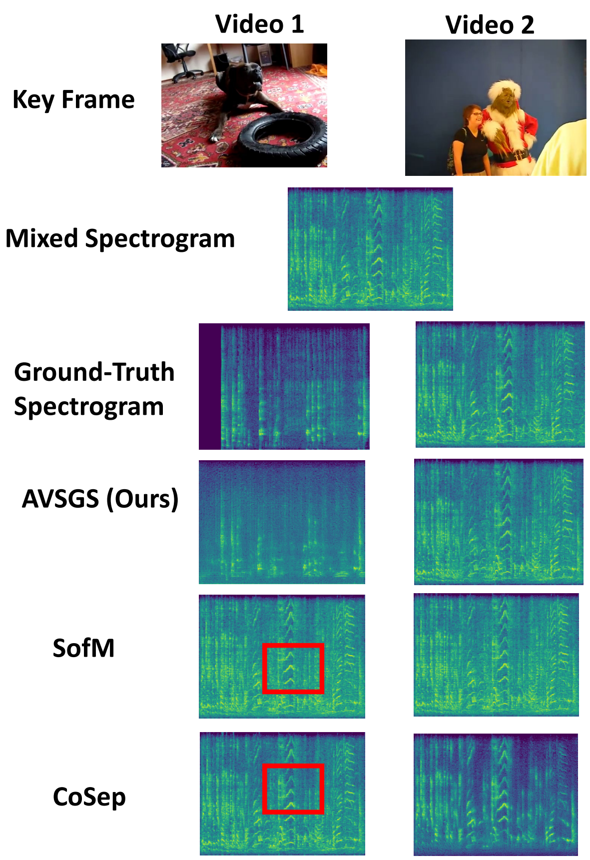

Figure 6. Qualitative separation results on a mixture of two ASIW videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

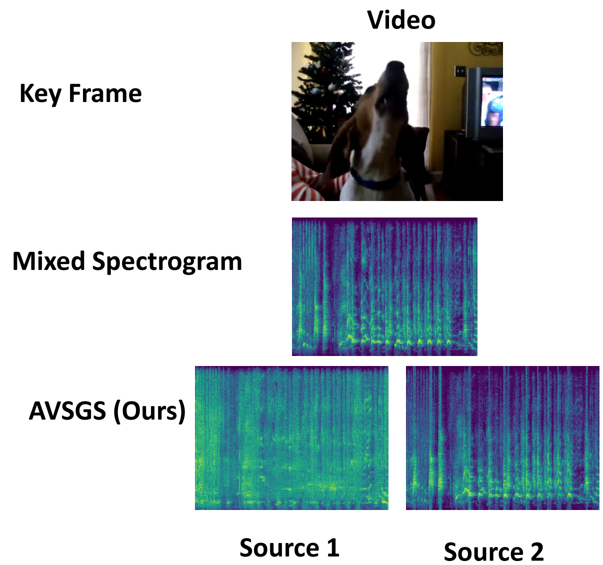

Figure 7. Qualitative separation results on a video with 2 sound sources for the ASIW videos. Sample key frame is shown for the videos are shown. The spectrogram of the separated audio is plotted.

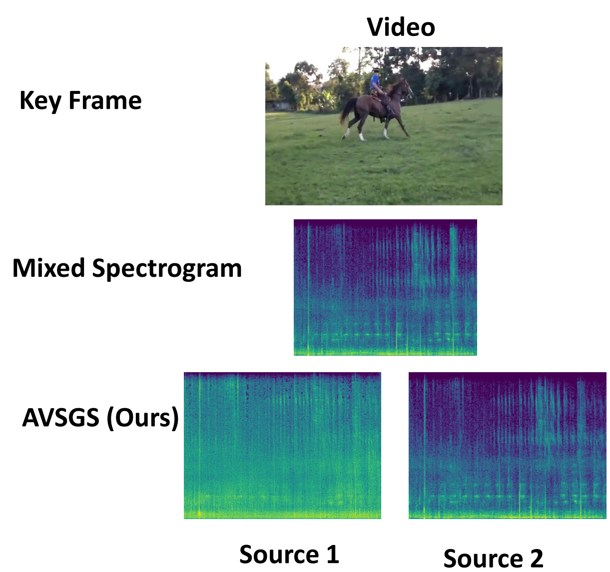

Figure 8. Qualitative separation results on a video with 2 sound sources for the ASIW videos. Sample key frame is shown for the videos are shown. The spectrogram of the separated audio is plotted.

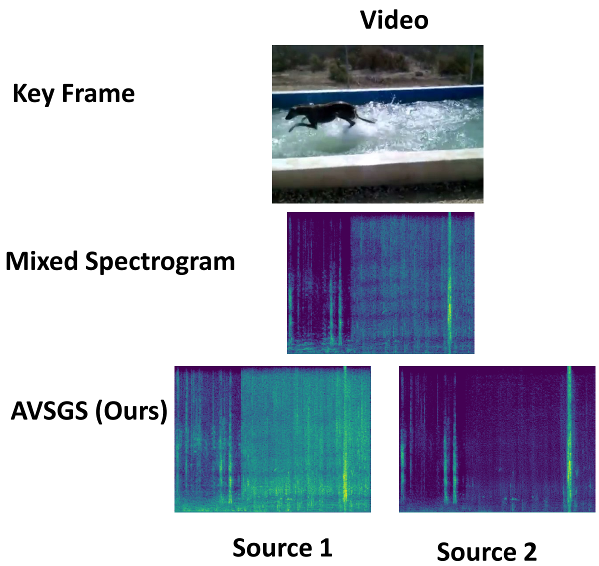

Figure 9. Qualitative separation results on a video with 2 sound sources for the ASIW videos. Sample key frame is shown for the videos are shown. The spectrogram of the separated audio is plotted.

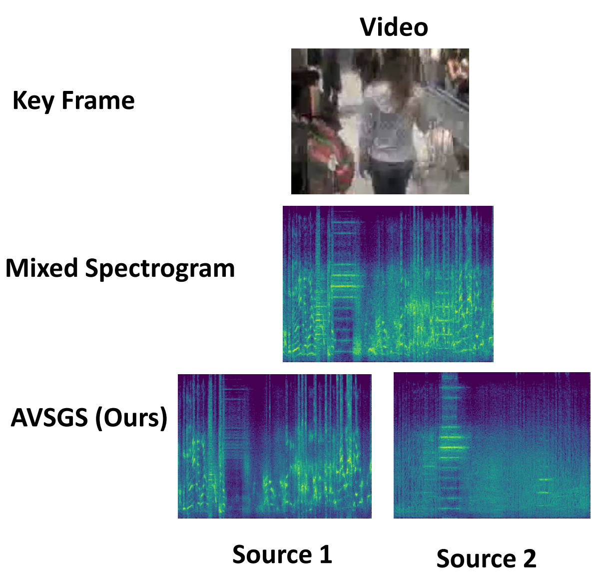

Figure 10. Qualitative separation results on a video with 2 sound sources for the ASIW videos. Sample key frame is shown for the videos are shown. The spectrogram of the separated audio is plotted.

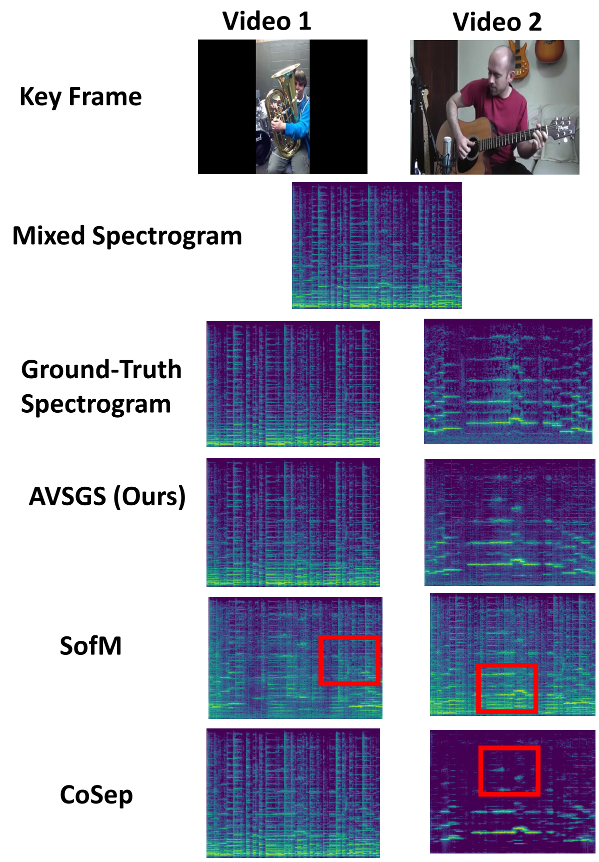

Figure 11. Qualitative separation results on a mixture of two MUSIC videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

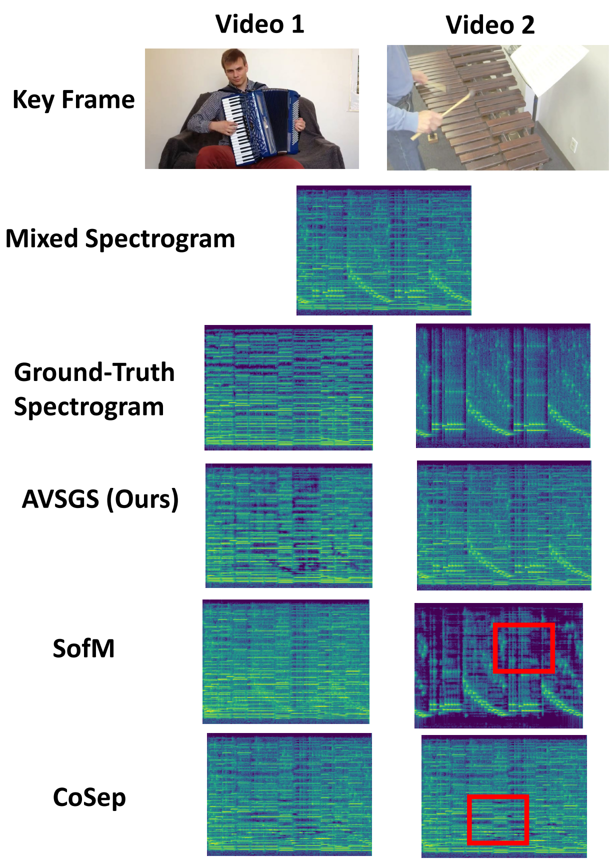

Figure 12. Qualitative separation results on a mixture of two MUSIC videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

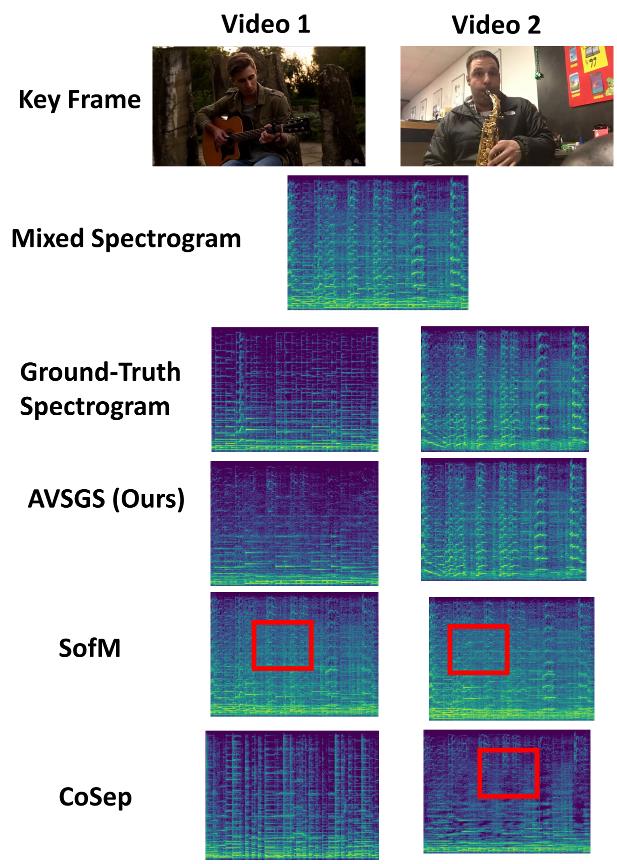

Figure 13. Qualitative separation results on a mixture of two MUSIC videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

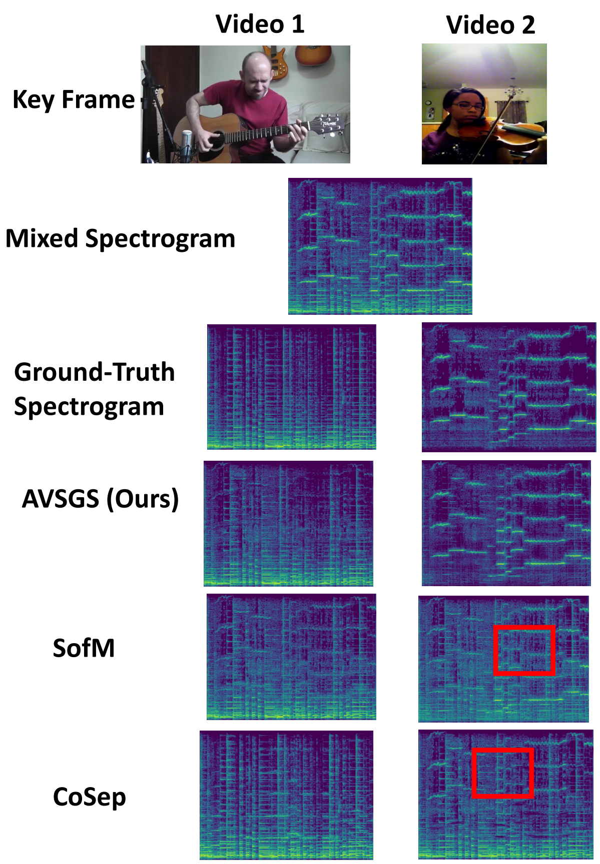

Figure 14. Qualitative separation results on a mixture of two MUSIC videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

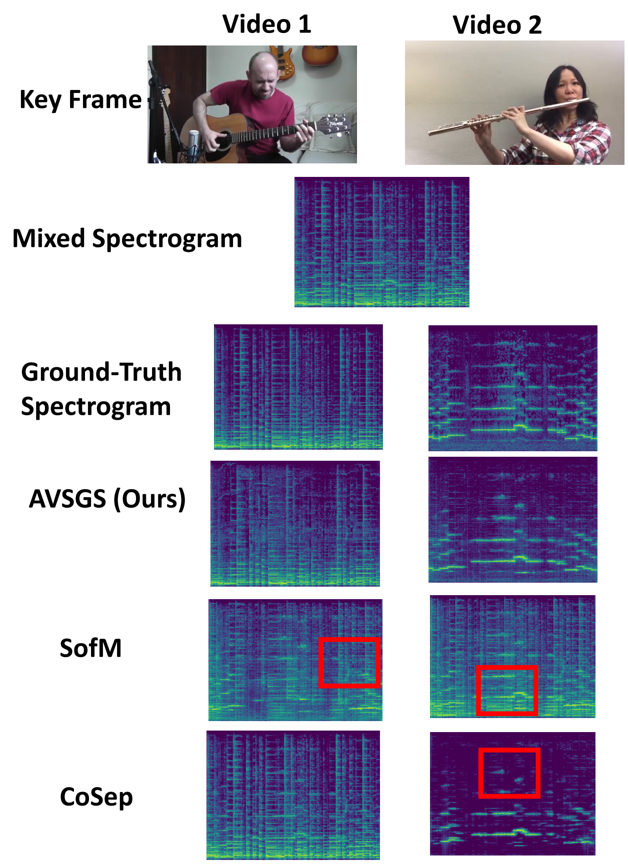

Figure 15. Qualitative separation results on a mixture of two MUSIC videos. Sample key frames for both videos are shown. The spectrogram of the mixed audio is plotted as well. Also shown are the separated spectrograms obtained by different methods. Red boxes indicate regions of high differences between ground truth and predicted spectrograms.

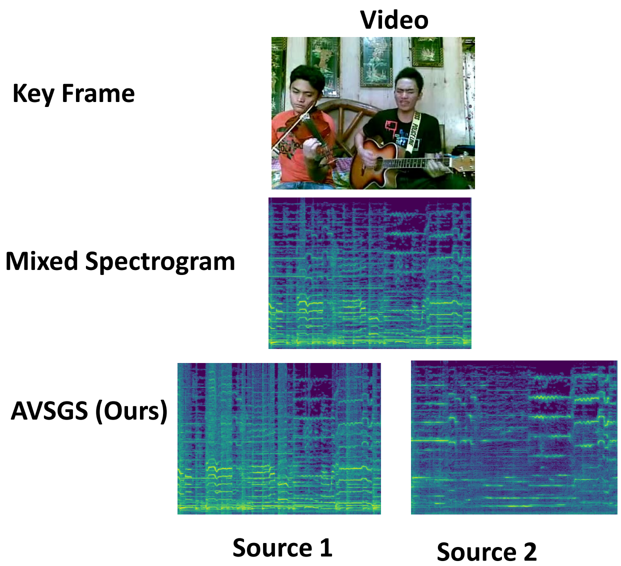

Figure 16. Qualitative separation results on a video with 2 sound sources for the MUSIC videos. Sample key frame is shown for the videos are shown. The spectrogram of the separated audio is plotted.

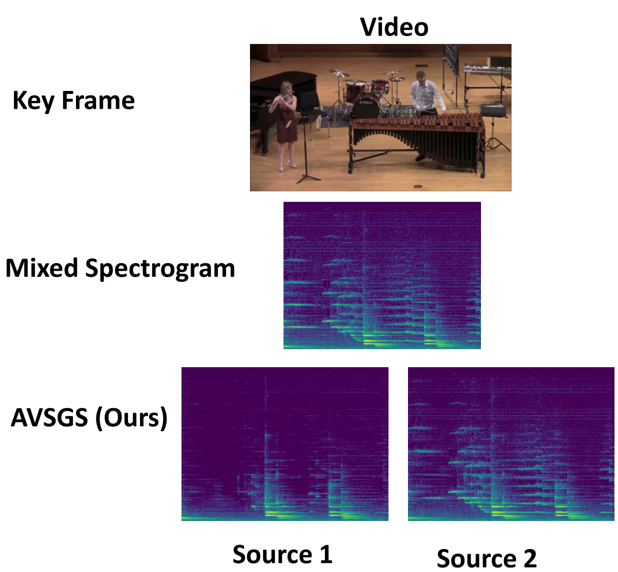

Figure 17. Qualitative separation results on a video with 2 sound sources for the MUSIC videos. Sample key frame is shown for the videos are shown. The spectrogram of the separated audio is plotted.

5. List of Auditory Words, Principal Objects, and Frequency in the ASIW Dataset

1. **babble:** baby/child/little girl 45
2. **babbling:** baby/child/little girl 8
3. **cry:** baby/child/little girl 1363
4. **crying:** baby/child/little girl 160
5. **fidget:** baby/child/little girl 9
6. **giggling:** baby/child/little girl 6
7. **jabbering:** baby/child/little girl 1
8. **singling:** baby/child/little girl 1
9. **sobbing:** baby/child/little girl 9
10. **sobs:** baby/child/little girl 12
11. **spitting:** baby/child/little girl 2
12. **chiming:** bell 44
13. **resonating:** bell 1
14. **rhythmically:** bell 47
15. **warning:** bell 59
16. **calling:** bird/birds/duck/ducks 10
17. **cheep:** bird/birds/duck/ducks 15
18. **chipping:** bird/birds/duck/ducks 4
19. **chirp:** bird/birds/duck/ducks 2274
20. **chirping:** bird/birds/duck/ducks 53
21. **flapping:** bird/birds/duck/ducks 14
22. **flutter:** bird/birds/duck/ducks 35
23. **gobbling:** bird/birds/duck/ducks 1
24. **quacking:** bird/birds/duck/ducks 104
25. **quaking:** bird/birds/duck/ducks 7
26. **squawk:** bird/birds/duck/ducks 73
27. **squawking:** bird/birds/duck/ducks 4
28. **vocalize:** bird/birds/duck/ducks 193
29. **whistling:** bird/birds/duck/ducks 100
30. **click:** camera 913
31. **donging:** clock/clocks/clock tower/alarm clocks 1
32. **locking:** clock/clocks/clock tower/alarm clocks 2
33. **tick:** clock/clocks/clock tower/alarm clocks 468
34. **ticking:** clock/clocks/clock tower/alarm clocks 187
35. **barking:** dog/dogs 591
36. **barks:** dog/dogs 1
37. **growl:** dog/dogs 305
38. **grumbling:** dog/dogs 2
39. **howl:** dog/dogs 127
40. **oinking:** dog/dogs 119
41. **panting:** dog/dogs 45
42. **playfully:** dog/dogs 10
43. **responding:** dog/dogs 3
44. **shakes:** dog/dogs 1
45. **whine:** dog/dogs 196
46. **yap:** dog/dogs 7
47. **emptying:** drain/toilet/toilet seat/toilet bowl 1
48. **flush:** drain/toilet/toilet seat/toilet bowl 824
49. **flushing:** drain/toilet/toilet seat/toilet bowl 13
50. **cantering:** horse/horses 1
51. **clop:** horse/horses 313
52. **clopping:** horse/horses 33
53. **galloping:** horse/horses 18
54. **neighs:** horse/horses 2
55. **oping:** horse/horses 1
56. **riding:** horse/horses 4
57. **oping:** horse/horses 1
58. **trotting:** horse/horses 12
59. **achoo:** man/woman/young man/people 1
60. **amplified:** man/woman/young man/people 9
61. **applaud:** man/woman/young man/people 289
62. **applauding:** man/woman/young man/people 22
63. **appreciatively:** man/woman/young man/people 1
64. **articulately:** man/woman/young man/people 1

65. **breathing:** man/woman/young man/people 132
66. **burp:** man/woman/young man/people 267
67. **celebrate:** man/woman/young man/people 2
68. **chant:** man/woman/young man/people 50
69. **chanting:** man/woman/young man/people 2
70. **cheer:** man/woman/young man/people 623
71. **cheering:** man/woman/young man/people 103
72. **chuckle:** man/woman/young man/people 98
73. **clapping:** man/woman/young man/people 40
74. **communicating:** man/woman/young man/people 6
75. **conversation:** man/woman/young man/people 158
76. **converse:** man/woman/young man/people 91
77. **coughing:** man/woman/young man/people 21
78. **coughs:** man/woman/young man/people 1
79. **crunching:** man/woman/young man/people 33
80. **curtly:** man/woman/young man/people 1
81. **dialog:** man/woman/young man/people 4
82. **echo:** man/woman/young man/people 141
83. **eruption:** man/woman/young man/people 3
84. **exhaling:** man/woman/young man/people 1
85. **falsetto:** man/woman/young man/people 1
86. **fighting:** man/woman/young man/people 3
87. **flicking:** man/woman/young man/people 1
88. **folding:** man/woman/young man/people 1
89. **forklift:** man/woman/young man/people 1
90. **gag:** man/woman/young man/people 6
91. **girlish:** man/woman/young man/people 1
92. **glee:** man/woman/young man/people 1
93. **hoots:** man/woman/young man/people 1
94. **indistinctly:** man/woman/young man/people 7
95. **inhale:** man/woman/young man/people 20
96. **kaboom:** man/woman/young man/people 1
97. **laugh:** man/woman/young man/people 3091
98. **laughing:** man/woman/young man/people 270
99. **manspaking:** man/woman/young man/people 1
100. **melody:** man/woman/young man/people 24
101. **moaning:** man/woman/young man/people 1
102. **monotone:** man/woman/young man/people 10
103. **murmur:** man/woman/young man/people 91
104. **narrating:** man/woman/young man/people 11
105. **playing:** man/woman/young man/people 123
106. **prancing:** man/woman/young man/people 1
107. **recording:** man/woman/young man/people 8
108. **reverberate:** man/woman/young man/people 9
109. **reverberating:** man/woman/young man/people 4
110. **screaming:** man/woman/young man/people 32
111. **scuffling:** man/woman/young man/people 4
112. **sigh:** man/woman/young man/people 39
113. **sighing:** man/woman/young man/people 2
114. **slurp:** man/woman/young man/people 10
115. **slurping:** man/woman/young man/people 1
116. **sneezing:** man/woman/young man/people 24
117. **sniffing:** man/woman/young man/people 7
118. **sniveling:** man/woman/young man/people 1
119. **snort:** man/woman/young man/people 65
120. **stuttering:** man/woman/young man/people 4
121. **subdued:** man/woman/young man/people 5
122. **thumping:** man/woman/young man/people 123
123. **thunderous:** man/woman/young man/people 5
124. **uproar:** man/woman/young man/people 4
125. **uproarious:** man/woman/young man/people 1
126. **uproariously:** man/woman/young man/people 1
127. **verbally:** man/woman/young man/people 2
128. **vigorously:** water/water tank/water bottle 17
129. **yelling:** man/woman/young man/people 74
130. **yodel:** man/woman/young man/people 1

131. **baaing:** sheep/goat/goats/chicken 114
132. **bleat:** sheep/goat/goats/chicken 583
133. **cackle:** sheep/goat/goats/chicken 13
134. **answering:** telephone 4
135. **ringing:** telephone 218
136. **chug:** train/trains/train car/train cars/passenger train/train engine 133
137. **sounding:** train/trains/train car/train cars/passenger train/train engine 8
138. **backing:** vehicle/car/cars/truck/trucks 2
139. **beeps:** vehicle/car/cars/truck/trucks 2
140. **brake:** vehicle/car/cars/truck/trucks 76
141. **braking:** vehicle/car/cars/truck/trucks 2
142. **breaks:** vehicle/car/cars/truck/trucks 1
143. **driving:** vehicle/car/cars/truck/trucks 25
144. **honk:** vehicle/car/cars/truck/trucks 584
145. **racing:** vehicle/car/cars/truck/trucks 50
146. **raggedly:** vehicle/car/cars/truck/trucks 1
147. **roving:** vehicle/car/cars/truck/trucks 2
148. **shifting:** vehicle/car/cars/truck/trucks 16
149. **silently:** vehicle/car/cars/truck/trucks 3
150. **skidding:** vehicle/car/cars/truck/trucks 15
151. **draining:** water/water tank/water bottle 1
152. **drip:** water/water tank/water bottle 106
153. **flowing:** water/water tank/water bottle 9
154. **gushing:** water/water tank/water bottle 1
155. **hisses:** water/water tank/water bottle 1
156. **jostling:** water/water tank/water bottle 2
157. **leaking:** water/water tank/water bottle 1
158. **pouring:** water/water tank/water bottle 14
159. **raining:** water/water tank/water bottle 6
160. **splashing:** water/water tank/water bottle 211
161. **splay:** water/water tank/water bottle 1
162. **trickling:** water/water tank/water bottle 22
163. **woosh:** water/water tank/water bottle 3
164. **audible:** background 22
165. **audibly:** background 1
166. **banging:** background 191
167. **beat:** background 53
168. **beatable:** background 1
169. **beating:** background 5
170. **beep:** background 910
171. **bellow:** background 2
172. **blast:** background 79
173. **blowing:** background 183
174. **boiling:** background 1
175. **bouncing:** background 6
176. **brushing:** background 4
177. **buffeting:** background 3
178. **bumble:** background 1
179. **burble:** background 31
180. **burbling:** background 1
181. **burning:** background 4
182. **bursting:** background 6
183. **buzzer:** background 13
184. **chang:** background 1
185. **chewing:** background 4
186. **chocking:** background 1
187. **choke:** background m 6
188. **churning:** background 3
189. **clacking:** background 173
190. **clang:** background 197
191. **clank:** background 597
192. **clanking:** background 334
193. **clattering:** background 39
194. **clinking:** background 76

195. **clumping:** background 1
196. **clunking:** background 8
197. **cluttering:** background 2
198. **cocking:** background 9
199. **collision:** background 3
200. **crack:** background 126
201. **cracking:** background 24
202. **cranking:** background 13
203. **crinkling:** background 103
204. **croak:** background 339
205. **croaking:** background 22
206. **crumpling:** background 63
207. **dabbling:** background 1
208. **deafen:** background 1
209. **dinging:** background 2
210. **drooping:** background 1
211. **explode:** background 53
212. **fainting:** background 1
213. **faintly:** background 253
214. **filing:** background 20
215. **firing:** background 35
216. **fizzing:** background 1
217. **flipping:** background 3
218. **fumbling:** background 1
219. **grinding:** background 34
220. **grunting:** background 28
221. **gulping :** background 2
222. **gusting:** background 4
223. **heaving:** background 1
224. **hoovering:** background 1
225. **hovering:** background 8
226. **humming:** background 639
227. **jarring:** background 1
228. **jumble:** background 1
229. **launching:** background 1
230. **licking:** background 1
231. **loudly:** background 1828
232. **mingle:** background 2
233. **mix:** background 22
234. **mixer:** background 1
235. **noise:** background 2529
236. **noisily:** background 13
237. **outburst:** background 3
238. **popping:** background 69
239. **pound:** background 24
240. **puffing:** background 2
241. **pulsing:** background 4
242. **rabbiting:** background 2
243. **ragged:** background 2
244. **raging:** background 1
245. **rapping:** background 6
246. **ratcheting:** background 5
247. **rattling:** background 113
248. **reeving:** background 1
249. **releasing:** background 20
250. **reloading:** background 5
251. **revving:** background 227
252. **rewinding:** background 1
253. **rhythm:** background 18
254. **ripping:** background 12
255. **roaring:** background 71
256. **rocking:** background 1
257. **roughly:** background 34
258. **rumbling:** background 72
259. **rustling:** background 543
260. **sanding:** background 20

261. **sawing:** background 30
262. **scowl:** background 1
263. **scrapping:** background 9
264. **shaking:** background 3
265. **sharpen:** background 3
266. **sharpening:** background 1
267. **shrill:** background 18
268. **slashing:** background 1
269. **slightly:** background 50
270. **smash:** background 6
271. **smashing:** background 4
272. **snare:** background 3
273. **snarling:** background 1
274. **sparking:** background 2
275. **splat:** background 7
276. **splattering:** background 1
277. **spraying:** background 70
278. **springing:** background 1
279. **spurt:** background 4
280. **squeaking:** background 61
281. **squealing:** background 60
282. **steadily:** background 76
283. **stitching:** background 7
284. **stretching:** background 1
285. **striking:** background 2
286. **suction:** background 5
287. **swarm:** background 40
288. **swishing:** background 21
289. **tapping:** background 192
290. **thud:** background 83
291. **thudding:** background 1
292. **thwacking:** background 3
293. **tinkling:** background 8
294. **trill:** background 3
295. **tumbling:** background 2
296. **typing:** background 129
297. **vibrantly:** background 1
298. **vibrate:** background 397
299. **vibrating:** background 66
300. **weirdly:** background 1
301. **whirring:** background 176
302. **whooshing:** background 123
303. **winding:** background 2
304. **wishing:** background 1
305. **zapping:** background 1
306. **zipping:** background 3

References

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[2] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proc. CVPR*, pages 10478–10487, 2020. 1

[3] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proc. ICCV*, pages 3879–3888, 2019. 1, 2

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*, pages 776–780, Mar. 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 2

[6] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proc. NAACL HLT*, pages 119–132, 2019. 1

[7] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 2

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1, 2

[9] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *Proc. ICML*, pages 3734–3743, June 2019. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2016. 2

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, pages 234–241. Springer, 2015. 2

[12] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *Proc. ICLR*, Apr. 2018. 2

[13] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph. (TOG)*, 38(5):1–12, 2019. 2

[14] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proc. ICCV*, pages 1735–1744, 2019. 1, 3

[15] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proc. ECCV*, pages 570–586, 2018. 1