INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Image-driven Audio-visual Universal Source Separation

Chenxing Li[1], *Ye Bai*[1], *Yang Wang*[1], *Feng Deng*[1], *Yuanyuan Zhao*[2], *Zhuo Zhang*[1], *Xiaorui Wang*[2]

[1] Kuaishou Technology Co., Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{lichenxing007}@gmail.com

Abstract

This paper introduces an image-driven audio-visual universal source separation (ID-USS) and proposes ID-USS-Conformer. ID-USS aims to separate a target source from the mixture based on the input image that is consistent with the target. Importantly, ID-USS only focuses on the sound made by the target in this image, not on the description of the target or the semantic information of the picture. In detail, ID-USS-Conformer mainly consists of an Efficient-b3-based visual branch and a Conformer-based audio branch. The visual branch extracts the visual clue of the target from the input image. After the audio branch fuses the visual features, ID-USS-Conformer separates the target source from the mixture. We launch an ID-USS dataset and verify the effectiveness of ID-USS-Conformer on it. The ID-USS-Conformer has achieved a 10.139 dB signal-to-distortion ratio improvement in the test set and outperformed the compared methods.

Index Terms: audio-visual source separation, universal source separation, image-driven target source separation

1. Introduction

Short videos have gradually become one of the ways for people to entertain and get in touch with the news. Audio understanding technology now in short videos mainly revolves around human voice, such as subtitle speech recognition, speaker recognition, etc. With the urgent need for understanding various audio events, it is necessary to separate and identify the acoustic event in short videos, such as the sounds of animals, vehicles, and natural events. We expect to edit the sounds of short videos, such as adding/deleting sound effects, build a large-scale sound library, and help the audio classification model to obtain better classification performance through separation technology.

This paper introduces image-driven audio-visual universal source separation (ID-USS) and proposes an ID-USS model to separate target audio. In detail, the object appearing in the image is identified as a detached target. The model uses images as cues to drive the model to separate the possible sounds of the target. ID-USS aims to separate universal acoustic events.

The single-channel multi-speaker speech separation has achieved remarkable results. Permutation invariant training (PIT) [1], CBLDNN-GAT [2], Conv-TasNet [3], and DPRNN [4] have successively achieved state-of-the-art performance. Multi-speaker separation mainly focuses on the separation between human voices rather than the separation of large-scale acoustic events in life. On this basis, Multi-task audio source separation (MTASS) [5–7] aims to separate speech, music, and noise/sound effects into three tracks at once. Universal sound separation [8–10] separates mixtures of arbitrary sounds of different types.

The popular separation methods can not meet our needs: (1) multi-speaker speech separation only focuses on human speech. (2) MTASS only separates speech, music, and noise. MTASS ignores the separation of acoustic events. Various sounds, such as closing doors, animal calls, and some annoying noises, are all classified as noise track signals. (3) USS achieves the separation of various acoustic events. However, it meets several challenges as PIT-based methods: an unknown number of sources in the mixture, permutation problem, and selection from multiple outputs. Since the USS model needs to fix the number of outputs in advance, USS often performs poorly when the number of acoustic events in the mixture exceeds the number of model outputs. Besides, when the number of outputs is large, this will increase the complexity of the PIT loss and training time.

Target-driven USS methods [11–13] adopt target clues to drive the audio model to perform the sound separation. This method effectively solves the problems encountered by the PIT-based USS methods. Paper [11] first uses the predictions of a sound classifier as embedding and then conditions on embedding to perform sound separation iteratively. The performance depends on the accuracy of the sound classifier. Sound selector [12] extracts the desired acoustic sound from the mixture, while a one-hot vector representing the class of interest is injected into the model. Sound selector [12] encodes 41 event sound sources, which makes it hard to support sound separation outside of classes. Class extending requires model retraining. Soundfilter [13] uses the same type of audio signal as a clue to drive USS. Due to the distribution change of audio, the selection of the reference has a significant influence on the performance. In practical applications, due to the complexity of audio visualization, it is hard to quickly find a matched reference audio.

Recently, text or image-driven separation models have been gradually proposed. Text or image-based methods show advantages over audio-based methods: (1) text and image can be easily recognized and acquired by humans; (2) text and image have low redundancy and strong expressive ability over audio. LASS [14] separates the target source from an audio mixture based on a natural language query of the target source. Similarly, text-driven Soundfilter [15] is first conditioned on arbitrary textual descriptions of sound or alternative audio and then separates corresponding sound. CO-SEPARATION [16] first detects target objects in video. The audio-visual separator network next takes a mixed audio signal and the detected object from its accompanying video as input and separates the sound responsible for the input object region. In descending order of average energy, MP-Net [17] separates sound from the recorded mixture based on a corresponding video. ConceptBeam [18] uses a concept specifier, such as an image or speech, to extract the speech of speakers speaking about a concept. Audioscope [19, 20] separates all audio associated with the video.

10.21437/Interspeech.2023-1309

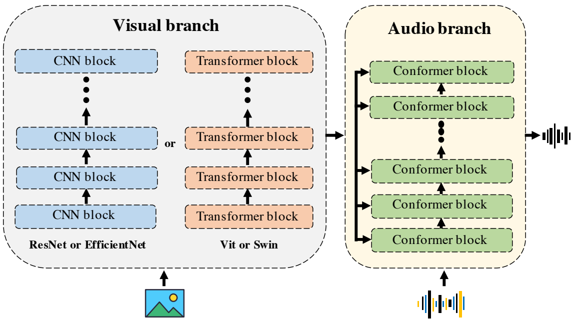

Figure 1: *The illustration of ID-USS-Conformer. Compared with various visual models, ID-USS-Conformer uses Efficient-b3 in the visual branch. The visual branch extracts target embedding from the image and injects it into the audio branch. The audio branch fuses visual embedding and separates the target.*

Text-driven separation methods learn target clues from text descriptions. The separation target of these methods is the semantic information or several acoustic events contained in the text. Image-based methods are language-independent, which shows an advantage over text-based methods. ConceptBeam [18] extracts the speech of speakers speaking about a concept, not the sound produced by the acoustic events represented in the image. CO-SEPARATION [16] and MP-Net [17] mainly focus on separating multi-sources driven by multi-source videos. The generalization and the detection accuracy of object detector may limit the performance of CO-SEPARATION [16]. Due to the out-of-date model structure, the performance of CO-SEPARATION [16] and MP-Net [17] is poor. Such scenarios are not aligned well with our goals.

In this paper, we propose an ID-USS-Conformer, which extracts event clue from the image and drive the separation model to separate the target source. In ID-USS-Conformer, the vision branch adopts the Efficient-b3 to encode the image into a vision embedding. A Conformer-based separation network is then used to separate the target source from the mixture conditioned on the vision embedding.

Based on COCO [21] and FSD50K [22], we build an open-source ID-USS dataset to verify the effectiveness of ID-USS-Conformer. ID-USS dataset includes common acoustic events in life. In this experiment, the visual pattern in the image only corresponds to a single acoustic event. If it is necessary to separate the sounds of multiple acoustic events, separation can be performed iteratively. Signal-to-distortion ratio improvement (SDRi) [23] is used to evaluate the performance. Experimental results show the effectiveness of the ID-USS-Conformer, which achieves 10.139 dB SDRi and outperforms several baselines.

2. System overview

The illustration of the proposed ID-USS-Conformer is depicted in Figure 1. ID-USS-Conformer consists of two parts: a visual branch and an audio branch. The visual branch extract target embedding lies in the image. The audio branch fuses the visual feature and separates the target source.

2.1. Visual branch

In the visual branch, the vision model extracts visual embedding of the target event. To obtain a better performance, the vision model is selected from pre-trained models. Two types of pre-trained models are selected: pre-trained models on imagenet classification and CLIP-based pre-trained models. The pre-trained models obtained through imagenet classification have better classification and discrimination power. CLIP-based pre-trained models take more semantic information and generalize well. For pre-trained models on imagenet classification, several classic models are selected in comparison, which are pre-trained ResNet-50 [24], Effecient-b3 [25], Vit [26], and Swin [27]. For CLIP-based pre-trained models, ResNet-50 in CLIP [28] and ResNet-50 in AudioCLIP [29] are selected.

Specifically, in pre-trained models, the output of the final convolutional layer prior to pooling is selected as the output of the visual model. The formulation of the visual branch is listed as:

$$E_{im} = ReLU(f(VM(Im))), \tag{1}$$

where Im is the input image. The visual model, VM, keeps the same structure as the pre-trained model and is initialized using pre-trained parameters. After passing through the visual model, visual embedding E_{im} is obtained via a feed-forward layer with 256 nodes and ReLU activation.

2.2. Audio branch

The audio branch fuses visual cues and drives the separation network to separate the target source. As the Conformer-based separation model has achieved superior results in continuous speech separation [30] and MTASS [7], a Conformer-based separation model is applied in the audio branch. The detailed structure of the Conformer block is shown in Figure 2.

In each Conformer block, the feature-wise Linearly modulated (FiLm) [31] layer transforms visual embedding into visual clues. Visual clues are added after the feed-forward network (FFN) module. The detailed formulation of FiLm is as follows:

$$FiLm(H, E_{im}) = g_1(E_{im})H + g_2(E_{im}), \tag{2}$$

where H represents the output of FFN in the Conformer block. g_1 and g_2 are feed-forward layers. E_{im} is the visual embedding obtained from the visual branch.

In this experiment, the separation model consists of 16 Conformer blocks. Each Conformer block consists of 4 attention heads, 256 attention dimensions, and 1024 FFN dimensions. In the convolution part, an additional squeeze-excitation layer [32] with $reductionRatio = 8$ is added. For the FiLm layer, feed-forward layers have 1024 dimensions.

2.3. The pipeline of ID-USS-Conformer

2.3.1. Model input

ID-USS-Conformer is conducted in the frequency domain, and spectral magnitude is selected as the input to the model. In detail, the mixture is first transformed by a short-time Fourier transform (STFT). For computing STFT, we use a 1024-sample window size and a 256-sample shift.

2.3.2. Audio-visual process

ID-USS-Conformer receives the image and spectral magnitude of the mixture and outputs the estimated spectral magnitude. The formula is as follows:

$$|\hat{S}(t,f)| = \textbf{ID-USS-Conformer}(|Y(t,f)|, Im), \tag{3}$$

where $|Y(t,f)|$ denotes the spectral magnitude of the mixture. Im is the input image, which draws the target event. $\hat{S}(t,f)$ represents the estimated spectral magnitude of the source.

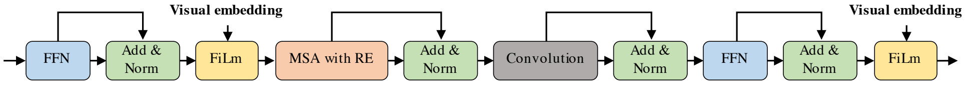

Figure 2: *The pipeline of Conformer block. After the FFN layer, The FiLm layer fuses acoustic features and the target visual embedding.*

2.3.3. Model output

ID-USS-Conformer outputs the spectral magnitude of the estimated source. After separation, we use inverse STFT (ISTFT) to restore the separated waveform. The phase of the mixture is used to restore the separated source.

2.4. Loss function

Two kinds of loss are applied: magnitude-based mean absolute error (MAE) loss and time domain-based SDR loss. MAE loss and SDR loss guide training in terms of energy and direction between the clean source and the estimated source. The detailed loss function is formulated as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{MAE} - \lambda \times \mathcal{L}_{SDR}, \\ where\ \mathcal{L}_{MAE} &= ||S - \hat{S}||, \\ \mathcal{L}_{SDR} &= SDR(s, \hat{s}), \end{aligned} \tag{4}$$

where λ is the scale parameter to balance each loss. S and $\hat{S}$ represent the spectral magnitude of the clean source and the estimated source. s and $\hat{s}$ are the ground truth waveform and the estimated waveform.

3. Datasets

3.1. Image data preparation

We choose the COCO dataset [21] for image data. COCO contains 90 types of events, a total of 123287 images. We cut out the image according to the location and annotation to ensure that there is only one event in the image. The area of the bounding part must be greater than 10000. After selection and cutting, 90 types of image data are formed, with a total of 207802 images.

3.2. Audio data preparation

The audio data comes from the FSD50k dataset [22]. FSD50k encompasses 200 sound classes and has a total of 108 hours of multi-labeled audio. Since the FSD50k may have multiple labels for a single audio, it may cause label inconsistencies and confusion between the image and audio. We use AudioSet Ontology [33] as the reference to select the audio that only contains one label. After the filtering, a total of 26525 pieces of audio are obtained, with a total of 102 events and 42 hours.

3.3. ID-USS dataset

For building the ID-USS dataset, we first pick out the matched types of audio-image pairs. A total of 20 types of audio-image pairs are selected. ID-USS dataset contains 20 types of paired audio-image data, which is selected as the target in the experiment. The left 82 types of audio act as interference in the ID-USS dataset.

In the ID-USS dataset, the data distribution of the target type is shown in Table 1. After splitting, the training set contains 103192 images with 20 target types, 9 hours of audio with 20 target types, and 26 hours of audio with 82 interference types. The development set has 4000 images with 20 target types, 0.5 hours of audio with 20 target types, and 1.5 hours of audio with 81 interference types. The test set has 6000 images with 20 target types, 1.12 hours of audio with 20 target types, and 3 hours of audio with 82 interference types. Importantly, there is no overlap among the training, development, and test sets for image and audio. The details of the ID-USS dataset, including the instructions and generation scripts, will be open-sourced soon.

Table 1: *The number of images and the audio duration of 20 target events in the ID-USS dataset.*

Target Event	Image count	Audio duration(s)
Person	68036	5187
Bottle	4970	1617
Dishes	3928	704
Car	3834	1430
Cat	3813	2566
Train	3704	3423
Bus	3622	2134
Motorcycle	3169	1505
Dog	3133	2405
Toilet	2672	2654
Microwave	2653	1519
Airplane	2403	1341
Sink	1535	2267
Bird	1392	820
Keyboard	1046	2507
Clock	980	2329
Cutlery	811	847
Cellphone	718	768
Skateboard	438	900
Scissors	335	1161

4. Experiments

4.1. Experimental setup

Model performance is evaluated by ID-USS data. During training, the image is randomly selected, and the label of this image is acted as the event target. For the mixture, the target audio is randomly selected but with the same event label. We randomly select 1-3 audios with different event labels and add them linearly to form background interference audio. The target audio and the background audio are mixed with a random SNR of -5 to 20 dB for each clip. Due to the uneven distribution of image count in different classes, a balanced sampling strategy [34] is used in training to ensure that each type of event can be fully trained. For a fair comparison, the data in the development set and the test set are mixed with the same rules as the training set. The selection of the target image, the target audio, and the interference audio are pre-fixed. The development set is used to guide the training process, and the test set is used to evaluate the performance. The SDR of the test set is 5.92 dB.

For the loss function, $\lambda = 1000$. The visual model is initialized with parameters from the pre-trained model, and the audio separation model starts from scratch. The audio and visual model update parameters along with the training. Models are trained with the AdamW optimizer with a 1e-3 learning rate.

The training schedule of self-attention-based models is a warm-up learning schedule with a linear decay, where the warm-up step is 12000. For convolution-based models, the learning rate is halved if the loss of the development set is not improved. Both models are trained with 200 epochs.

4.2. Baselines

For visual models, pre-trained models on imagenet classification[1] (ResNet-50 [24], Efficient-b3 [25], Vit [26], and Swin [27]) and CLIP-based pre-trained models (CLIP-ResNet-50[2] [28] and AudioCLIP-ResNet-50[3] [29]) are applied. During this process, the separation model adopts the Conformer structure.

For audio separation models, ResUNet[4] [14] and Soundfilter [15] are first selected, as they show superior performance in text-driven-based separation. We directly replace the text features with the visual features. We also reproduce the ConceptBeam [18] and CO-SEPARATION[5] [16] for comparison. In CO-SEPARATION, we omit the object detector and only use an audio-visual separator because the image in this experiment directly contains the target object. Conv-TasNet[6] [3] is also added for comparison.

4.3. Analysis on visual branch

We compare two types of pre-trained models to select the effective model in the visual branch. Conformer-based audio separation model is fixed in this comparison. Experimental results are listed in Table 2. For pre-trained models on imagenet classification, The Efficient-b3-based model performs better than the ResNet-50-based. This is due to the performance gap in image classification. Efficient-b3 can extract better discriminate features from the image. The Efficient-b3-based model performs similarly to Vit-based and Swin-based models. As the ID-USS dataset only consists of 20 types of images, the model may reach an upper limit on this dataset.

Although CLIP-based ResNet-50 and imagenet classification-based ResNet-50 use the same network structure, the information contained in them is different. CLIP and AudioCLIP pay more attention to semantic information and have strong generalization abilities. Imagenet classification-based ResNet-50 has better image classification performance. CLIP-ResNet-50 performs similarly to imagenet classification-based ResNet-50. AudioCLIP-ResNet-50 performs worse than CLIP-ResNet-50 and imagenet classification-based ResNet-50. AudioCLIP-ResNet-50 is finetuned while training AudioCLIP, resulting in a decline in the image classification performance. This demonstrates that the classification ability of the vision model contributes more to the separation performance in our task. For effectiveness and efficiency, ID-USS-Conformer selects Efficient-b3 as the visual model to extract visual embedding.

4.4. Results analysis

On the basis of Efficient-b3, we evaluate the impact of different separation models on the separation performance. We also compare ID-USS-Conformer with the popular models. The results are listed in Table 3. The Conformer-based model (ID-USS-Conformer) outperforms the ResUNet, Soundfilter, and Conv-Tasnet-based methods. Conformer is good at capturing long-term dependencies and local correlations, which may contribute to performance improvement. Besides, ConceptBeam adopts VGG16 [35, 36] in the visual branch and a recurrent network-based separation model in the audio branch. The visual model in ConceptBeam aims to bridge spoken audio with semantically relevant images. CO-SEPARATION applies imagenet pre-trained ResNet-18 [24] in the visual branch and a convolutional network-based separation model in the audio branch. ID-USS-Conformer performs better, resulting from the better modeling ability of both the visual and audio branches. ID-USS-Conformer achieves the best on the test set, which is 10.139 dB SDRi.

Table 2: *Parameters and SDRi for the comparison models.*

Visual model	Audio model	Params (M)	SDRi (dB)
ResNet-50 [24]	Conformer	56.42	8.682
Efficient-b3 [25]	Conformer	43.57	10.139
Vit [26]	Conformer	116.94	10.120
Swin [27]	Conformer	118.08	10.233
CLIP-ResNet-50 [28]	Conformer	54.61	8.736
AudioCLIP-ResNet-50 [29]	Conformer	54.61	8.254

Table 3: *Parameters and SDRi for the comparison models.*

Visual model	Audio model	Params (M)	SDRi (dB)
Efficient-b3	Conv-TasNet [3]	34.53	8.686
Efficient-b3	Soundfilter [15]	30.40	8.738
Efficient-b3	ResUNet [14]	76.49	9.431
CO-SEPARATION [16]		57.20	7.579
ConceptBeam [18]		72.82	7.815
ID-USS-Conformer		43.57	**10.139**

5. Discussions and conclusions

In this paper, the proposed ID-USS-Conformer separates a target source from the mixture driven by the visual target that lies in the input image. In detail, Efficient-b3 is selected in the visual branch to process image and generate target visual embedding. Conformer-based separation model first fuses visual embedding and then separates the target source. We also propose an ID-USS dataset to evaluate the feasibility of ID-USS and the effectiveness of ID-USS-Conformer.

From experimental results, the image classification ability of the vision model has a positive impact on final separation performance. For the audio branch, Conformer performs well. This is due to the fact that the Conformer can better model global dependencies. ID-USS-Conformer achieves 10.139 dB SDRi, which also outperforms several baselines. The codes, the pre-trained models, and the analyses will be released soon.

In practice, ID-USS can be used to better edit short videos, such as precisely removing certain sound effects. Besides, by applying ID-USS-Conformer as the front end, the performance of the audio classification system can be improved. Due to the limitation of the ID-USS dataset, the experiments only focus on separating 20 types of targets. We pay more attention to verifying the feasibility of ID-USS and the effectiveness of ID-USS-Conformer. In the future, the ID-USS dataset will be expanded. The class number of target/interference and the quantity of image and audio data will be expanded.

[1] https://github.com/huggingface/pytorch-image-models

[2] https://github.com/openai/CLIP

[3] https://github.com/AndreyGuzhov/AudioCLIP

[4] https://github.com/liuxubo717/LASS

[5] https://github.com/rhgao/co-separation

[6] https://github.com/naplab/Conv-TasNet

6. References

[1] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[2] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, “Cbldnn-based speaker-independent speech separation via generative adversarial training,” in *IEEE ICASSP*, 2018, pp. 711–715.

[3] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation,” in *IEEE ICASSP*, 2020, pp. 46–50.

[5] L. Zhang, C. Li, F. Deng, and X. Wang, “Multi-task audio source separation,” in *IEEE ASRU*, 2021, pp. 671–678.

[6] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, “The cocktail fork problem: Three-stem audio separation for real-world soundtracks,” in *IEEE ICASSP*, 2022, pp. 526–530.

[7] C. Li, Y. Wang, F. Deng, X. Wang, and Z. Wang, “Ead-conformer: A conformer-based encoder-attention-decoder-network for multi-task audio source separation,” in *IEEE ICASSP*, 2022, pp. 521–525.

[8] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, “Universal sound separation,” in *IEEE WASPAA*, pp. 175–179.

[9] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, “What’s all the fuss about free universal sound separation data?” in *IEEE ICASSP*, 2021, pp. 186–190.

[10] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, “Compute and memory efficient universal sound source separation,” *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[11] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, “Improving universal sound separation using sound classification,” in *IEEE ICASSP*, 2020, pp. 96–100.

[12] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, “Listen to what you want: Neural network-based universal sound selector,” in *Interspeech*, 2020, pp. 1441–1445.

[13] B. Gfeller, D. Roblek, and M. Tagliasacchi, “One-shot conditional audio filtering of arbitrary sounds,” in *IEEE ICASSP*, 2021, pp. 501–505.

[14] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, “Separate what you describe: language-queried audio source separation,” in *Interspeech*, 2022.

[15] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi, “Text-driven separation of arbitrary sounds,” in *Interspeech*, 2022.

[16] R. Gao and K. Grauman, “Co-separating sounds of visual objects,” in *IEEE/CVF ICCV*, 2019, pp. 3879–3888.

[17] X. Xu, B. Dai, and D. Lin, “Recursive visual sound separation using minus-plus net,” in *IEEE/CVF ICCV*, 2019, pp. 882–891.

[18] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki, D. Takeuchi, D. Niizumi, A. Kimura, N. Harada, and K. Kashino, “Conceptbeam: Concept driven target speech extraction,” in *ACM MM*, 2022, pp. 4252–4260.

[19] E. Tzinis, S. Wisdom, A. Jansen, S. Hershey, T. Remez, D. Ellis, and J. R. Hershey, “Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds,” in *ICLR*, 2021.

[20] E. Tzinis, S. Wisdom, T. Remez, and J. R. Hershey, “Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation,” in *ECCV*. Springer, 2022, pp. 368–385.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *ECCV*. Springer, 2014, pp. 740–755.

[22] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “Fsd50k: an open dataset of human-labeled sound events,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[23] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE CVPR*, 2016, pp. 770–778.

[25] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *ICML*. PMLR, 2019, pp. 6105–6114.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *ICLR*, 2021.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *IEEE/CVF ICCV*, 2021, pp. 10 012–10 022.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*. PMLR, 2021, pp. 8748–8763.

[29] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “Audioclip: Extending clip to image, text and audio,” in *IEEE ICASSP*, 2022, pp. 976–980.

[30] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, “Continuous speech separation with conformer,” in *IEEE ICASSP*, 2021, pp. 5749–5753.

[31] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *AAAI*, vol. 32, no. 1, 2018.

[32] A. G. Roy, N. Navab, and C. Wachinger, “Concurrent spatial and channel ‘squeeze & excitation’in fully convolutional networks,” in *MICCAI*. Springer, 2018, pp. 421–429.

[33] J. F. Gemmeke, D. P. 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 *IEEE ICASSP*, 2017, pp. 776–780.

[34] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.

[36] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *ECCV*, 2018, pp. 649–665.