V-SlowFast Network for Efficient Visual Sound Separation

Lingyu Zhu
Tampere University, Finland
lingyu.zhu@tuni.fi

Esa Rahtu
Tampere University, Finland
esa.rahtu@tuni.fi

Abstract

The objective of this paper is to perform visual sound separation: i) we study visual sound separation on spectrograms of different temporal resolutions; ii) we propose a new light yet efficient three-stream framework ***V-SlowFast*** *that operates on Visual frame,* ***Slow*** *spectrogram, and* ***Fast*** *spectrogram. The* ***Slow*** *spectrogram captures the coarse temporal resolution while the* ***Fast*** *spectrogram contains the fine-grained temporal resolution; iii) we introduce two contrastive objectives to encourage the network to learn discriminative visual features for separating sounds; iv) we propose an audio-visual global attention module for audio and visual feature fusion; v) the introduced* ***V-SlowFast*** *model outperforms previous state-of-the-art in single-frame based visual sound separation on small- and large-scale datasets: MUSIC-21, AVE, and VGG-Sound. We also propose a small* ***V-SlowFast*** *architecture variant, which achieves 74.2% reduction in the number of model parameters and 81.4% reduction in GMACs compared to the previous multi-stage models. Project page: https://ly-zhu.github.io/V-SlowFast*

1. Introduction

Sound source separation aims at extracting the target source from a given audio mixture. The audio-based source separation task [28, 49, 55, 16] has been extensively studied in the audio processing community. However, the task remains challenging due to the underdetermined nature of source separation problem. The cocktail party problem [29, 17] is a well known example, where one attempts to follow one of the discussions while multiple people are talking simultaneously.

Recent works [64, 17, 63, 25, 65, 66, 21, 67] have started to exploit visual information (e.g. talking face, playing instruments) to solve the sound separation task. For instance, visual cues like object categories or movements can be used to facilitate the source separation problem. While visual motions may be important under certain circumstances (e.g. separating similar type of sources), the single visual frame based approaches have demonstrated surprisingly well performance in [64, 65, 66]. In this paper, we focus on improving the single visual frame based sound separation.

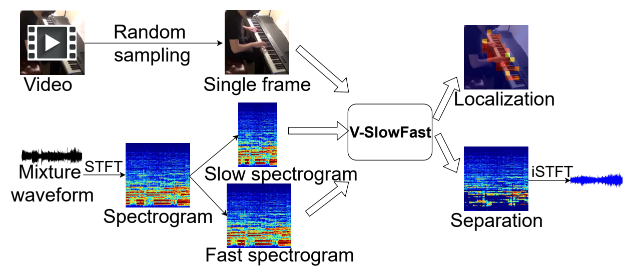

Figure 1. The **V-SlowFast** network operates on a visual frame (extracted from video), slow (low temporal resolution), and fast spectrogram (high temporal resolution). Eventually, the V-SlowFast model can efficiently separate and localize sound sources.

Natural sounds have wide range of rhythms. For example, slow attacks or fast tweaks occur fairly frequently when someone plays an instrument. In order to gain a new perspective on perceiving natural sounds in the sound separation task, we implement a system to treat the slow attacks and fast tweaks separately. The concept of slow-fast networks have shown impressive success in video [18, 59] and audio [36] recognition tasks, which operate on two streams of video frames or audio spectrograms with different sampling rates. Differently, we propose a novel three-stream framework **V-SlowFast** (Figure 1) for the visually guided sound separation task: **V**ision, **Slow**, and **Fast** pathway operating on visual frame, slow spectrogram, and fast spectrogram, respectively. The **Slow** spectrogram pathway has coarse temporal resolution (low sampling rate) while the **Fast** spectrogram pathway operates at fine-grained temporal resolution (high sampling rate). Moreover, we apply the concept of contrastive learning to the vision network for gaining discriminative semantic representations, which provide categorical cues (e.g. instrument type) for separating sounds and localizing sounding sources. Furthermore, we introduce an audio-visual global attention module (AVGA) to fuse the audio and visual features for making the model concentrate on the target sound source by leveraging corresponding global visual attention. Next, we upsample the global attended spectrum features to predict a mask for separating each component audio from mixture.

Multi-stage architectures [61, 65, 67] have shown good performance on visual source separation. However, these models tend to be large with high computational costs. We examine multiple options based on combinations of different spectrogram pathways (on different temporal resolutions) and different network architecture variants. To discover this, the **V-SlowFast** operates on spectrograms in multiple temporal resolutions. This is in contrast to previous works (e.g. [61, 65]), where the separation is done only based on the spectrogram with full temporal resolution at each stage. On the one hand, we show that the **V-SlowFast** network can greatly improve the sound separation performance (SDR: 10.89) over the recent single visual frame based multi-stage system [65] (SDR: 9.50) and recursive method [61] (SDR: 9.15). On the other hand, we also propose a small **V-SlowFast** architecture variant, which contains only 15M parameters and consumes 0.84 GMACs [41] for achieving similar result as previous multi-stage and recursive models (e.g. 58M parameters and 4$\sim$5 GMACs).

2. Related Work

Audio-Visual Learning Audio-visual learning combines signals from different modalities: audio and vision. Recent works [5, 3, 4] associate the learnt audio and visual embeddings by leveraging their correspondence. Synchronization based cross-modal approaches [45, 37, 15, 67] are proposed for visual representation learning. Another interesting task is to localize objects that sound [64, 4, 54, 65, 67, 11], where the goal is to pinpoint audio sources from the visual data. Other interesting works study audio-visual action recognition [35, 38, 26, 58], audio-visual navigation [22, 10, 9], talking head synthesis [56], spatial audio from video [43, 24, 62, 42], and visual-to-auditory [33, 20].

Visual Sound Separation Early work [6] performs audio-visual sound attribution by leveraging the tight associations between audio and visual onset signal. Recently, Zhao *et al.* [64, 63] proposed pioneering works to utilize appearance and motion cues for separating sound sources. Gao *et al.* [25, 23] studied to use object detection to facilitate source separation. Xu *et al.* [61] proposed a recursive model for separating sounds. Zhu *et al.* [65] further improved the models by utilizing visual cues of all the opponent sources. Gan *et al.* [21] associated keypoint-based body and finger movements with audio signals to separate sound sources. Owens *et al.* [45] and Zhu *et al.* [67] proposed synchronization based approaches for source separation. These works demonstrated how semantic appearances and motions could be utilized for sound separation. However, these works solely use full resolution spectrogram, which often leads to unnecessarily complex models.

Self-Supervised Contrastive Learning Contrastive learning leverages multiple perspectives of the data to learn discriminative features. It has been actively studied recently for images [44, 7, 66, 31, 14, 53], videos [57, 60, 34], text [52, 1], optical-flow [30, 53], and audio-video [46, 47, 2, 39, 67, 40]. In relation to previous efforts, our work studies two visual feature based contrastive objectives to obtain discriminative visual representation for visual sound separation and localization.

SlowFast Networks There is a classical branch of works focusing on the two-stream methods [51, 19, 8], which exploit two different stream modalities (e.g. RGB images and flow). Recently, Feichtenhofer *et al.* [18] introduced a SlowFast network, which contains two pathways separately working at low and high framerates for video recognition. Similarly, Kazakos *et al.* [36] proposed a two-stream convolutional network for audio recognition, that operates on low and high time-frequency spectrogram inputs. Xiao *et al.* [59] proposed slow and fast visual pathways that are integrated with a faster audio pathway to model vision and sound for video recognition. Inspired from previous research in multimodel and multi-resolution models, we propose a novel three-stream framework **V-SlowFast** for the visual sound separation task, which operates on visual frame and spectrograms of slow and fast sampling rates.

3. Approach

In this section, we first give a brief overview of our system (Sec. 3.1). Then we propose a novel **V-SlowFast** network for visual sound separation, that associates **V**ision (Sec. 3.2), spectrogram of **Slow** sampling rate (Sec. 3.3), and spectrogram of **Fast** sampling rate (Sec. 3.4). Finally, we present our learning objective in Section 3.5.

3.1. Overview

The goal of the visual sound separation is to extract the component audio that corresponds to the sound source in the given visual frame. Figure 2 illustrates the overall architecture of the proposed **V-SlowFast** network, which contains four components: vision network, audio-visual global attention module, slow spectrogram network, and fast spectrogram residual network. The vision network randomly extracts a single frame from the input video sequence and encodes it into a feature vector. To enhance the discrimination between semantic categories, we randomly sample an additional visual frame from a same (positive) or different (negative) category video to make contrastive pairs during the training procedure. We apply two visual contrastive objectives (embedding and localization) to the contrastive pairs along the vision network. The audio-visual global attention module fuses the visual embedding with sound fea-

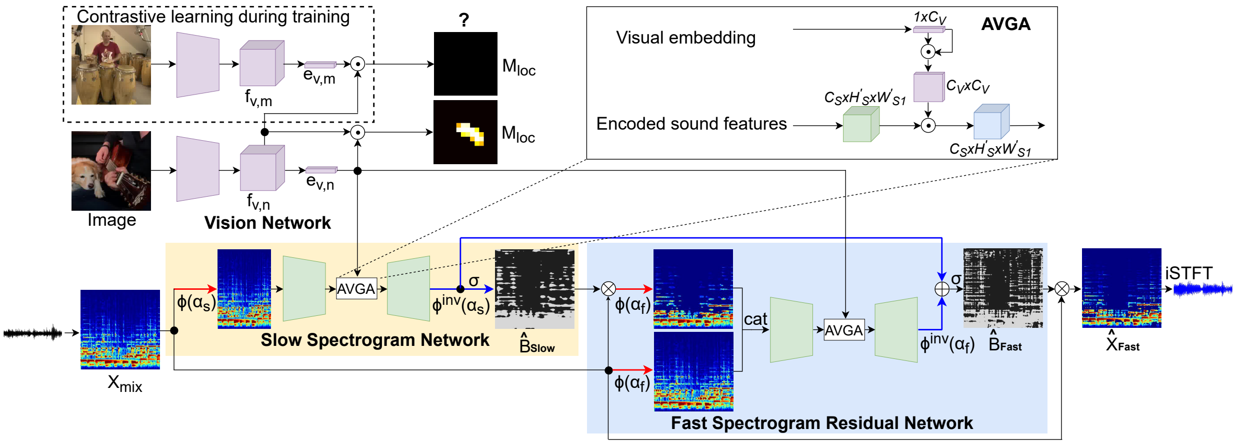

Figure 2. The overview of the proposed V-SlowFast framework for visual sound separation. It contains four components: vision network, audio-visual global attention (AVGA) module, slow spectrogram network, and fast spectrogram residual network. The slow and fast network operate on low and high temporal resolution spectrograms respectively. Note that the contrastive learning (dashed block) is only applied during training.

tures. The slow spectrogram network performs source separation at the coarse time scale (low sampling rate) using appearance features. The obtained result and the original mixture are further passed to the fast spectrogram residual network, which refines the source separation using spectrogram with higher temporal resolution (high sampling rate). The following sections provide further details of the system components and the learning objectives.

3.2. Vision Network

The Vision Network V receives a randomly sampled frame I from the input video and applies a dilated Res18-2D [32] or MV2 [50] to obtain a semantic representation e_v. More specifically, given an input RGB image $I \in \mathbb{R}^{3\times H_V\times W_V}$, the Vision Network produces feature maps $f_v \in \mathbb{R}^{C_V\times H_V^{'}\times W_V^{'}}$. These are passed to a spatial average pooling layer to obtain visual embedding $e_v \in \mathbb{R}^{1\times C_V}$,

$$f_v = V(I), \quad e_v = spatial\_pool(f_v), \tag{1}$$

where the C_V denotes the dimension of visual features. $H_V^{'} = H_V/16$ and $W_V^{'} = W_V/16$.

Visual Contrastive Learning Objectives We introduce two visual contrastive learning objectives: $\mathcal{L}_e(m,n,y)$ (embedding) and $\mathcal{L}_M(m,n,y)$ (localization) to the vision network. For the corresponding visual frame of each source n, we randomly sample an additional visual frame from a same or different category video m to form contrastive learning pair. $y = 0$ (negative) if m,n are of different type and 1 (positive) otherwise. We formulate the visual embedding contrastive learning objective $\mathcal{L}_e(m,n,y)$ to enforce large visual embedding distances between negative pair and small distances for positive pair. The inner product between the visual features $f_{v,n}$ and visual embedding $e_{v,m}$ yields a sounding source location mask $M_{loc}(m,n)$. Therefore, we define a localization contrastive objective $\mathcal{L}_M(m,n,y)$, with a binary cross entropy (BCE) loss between the location mask $M_{loc}(m,n)$ and y, to enforce empty localization mask between negative pair and non-empty mask of sounding objects for positive pair. More specifically,

$$\begin{aligned} dist(m,n) =& \sum \left(e_{v,m} - e_{v,n}\right)^2, \\ \mathcal{L}_e(m,n,y) =& \frac{1}{2} y \cdot dist(m,n) + \frac{1}{2}(1-y) \cdot \\ & max\{0, margin - \sqrt{dist(m,n) + e^{-9}}\}^2, \\ M_{loc}(m,n) =& pool\big(\sigma(e_{v,m} \odot f_{v,n})\big), \\ \mathcal{L}_M(m,n,y) =& BCE(M_{loc}(m,n), y), \\ \mathcal{L}_{contrast} =& r_1 \cdot \mathcal{L}_e(m,n,y) + r_2 \cdot \mathcal{L}_M(m,n,y), \end{aligned} \tag{2}$$

where the $dist(m,n)$ indicates the visual embedding distance between source m and n. We adopt $margin = 1.0$. A scalar product between the semantic embedding $e_{v,m}$ and the visual features $f_{v,n}$ results in a location mask $M_{loc}(m,n)$. σ and $pool$ represent the $sigmoid$ and max_pool operation, respectively. r_1=r_2=0.1 control the contribution of each objective factor.

3.3. Slow Spectrogram Network

We adopt an encoder-decoder style architecture of U-Net [48] or DeepLabV3Plus [13] for the slow spectrogram network. The U-Net consists of 7 down- and 7 up-convolutional layers with skip connection followed by a BatchNorm layer and a Leaky ReLU. MobileNetV2 (MV2) [50] is adopted as the backbone of the DeepLabV3Plus. The input of the slow spectrogram network is a 2D frequency-time spectrogram of mixture sound and the output is a same-size binary spectrogram mask.

Encoder The original audio mixture waveform is converted to a spectrogram presentation $X_{mix} \in \mathbb{R}^{1 \times H_S \times W_S}$ using Short-time Fourier Transform (STFT). We downsample the 2D frequency-time spectrogram X_{mix}, with a downsampling rate of α_s along the temporal dimension, as $\phi(X_{mix}, \alpha_s)$. The encoder $Slow^E$ takes the mixture spectrogram $\phi(X_{mix}, \alpha_s)$ as input and produces a feature representation $f_{Slow^E, mix, \alpha_s}$, as follows,

$$f_{Slow^E, mix, \alpha_s} = Slow^E\big(\phi(X_{mix}, \alpha_s)\big), \quad \alpha_s > 1, \tag{3}$$

where the ϕ represents the temporal downsampling operation, α_s indicates the downsampling rate for the slow spectrogram network. $\phi(X_{mix}, \alpha_s) \in \mathbb{R}^{1 \times H_S \times W_{S1}}$ and $f_{Slow^E, mix, \alpha_s} \in \mathbb{R}^{C_S \times H_S^{'} \times W_{S1}^{'}}$. C_S, H_S and W_S denote the dimension of sound features, and frequency-time bases of the sound spectrogram, respectively. The $W_{S1} = W_S/\alpha_s$, $H_S^{'} = H_S/16$ and $W_{S1}^{'} = W_{S1}/16$.

Audio-Visual Global Attention module Previous works [45, 63, 21, 67] have exploited the early fusion of the visual and audio features. However, these methods are mainly designed for fusing the visual motions into the sound features. In this paper, we propose an audio-visual global attention module (AVGA) in Figure 2 to fuse the semantic embedding e_v into the middle part of the sound spectrogram network for making the model concentrate on the target source by leveraging corresponding global visual attention. In order to keep the channel and spatial dimension of the sound features unchanged, we adopt a self attention to the semantic embedding before fusing with the sound features.

$$f_{Slow^{E,V}, n, \alpha_s} = e_{v,n}^T \odot e_{v,n} \odot f_{Slow^E, mix, \alpha_s}, \tag{4}$$

where the $\odot$ represents scalar product. The visual features $e_{v,n}$ of n-th source first have a self-attention with its own transposed embedding, then multiply with the encoded slow mixture sound features for providing global categorical attention. $f_{Slow^{E,V}, n, \alpha_s} \in \mathbb{R}^{C_S \times H_S^{'} \times W_{S1}^{'}}$ denotes the n-th global visual-attended sound features. Note that C_S equals to the C_V in Section 3.2.

Decoder The global visual-attended sound features are passed to the up-convolutional decoder ($Slow^D$) to produce a spectrum mask $f_{Slow^{D,V}, n, \alpha_s} \in \mathbb{R}^{1 \times H_S \times W_{S1}}$, which has the same-size as the temporally downsampled input spectrogram $\phi(X_{mix}, \alpha_s)$. In order to make the output to have the same resolution as the original spectrogram, we apply ϕ^{inv} (inverse ϕ) operation with the stride value of α_s. Following sigmoid operation results in a binary mask $\hat{B}_{Slow,n,1}$,

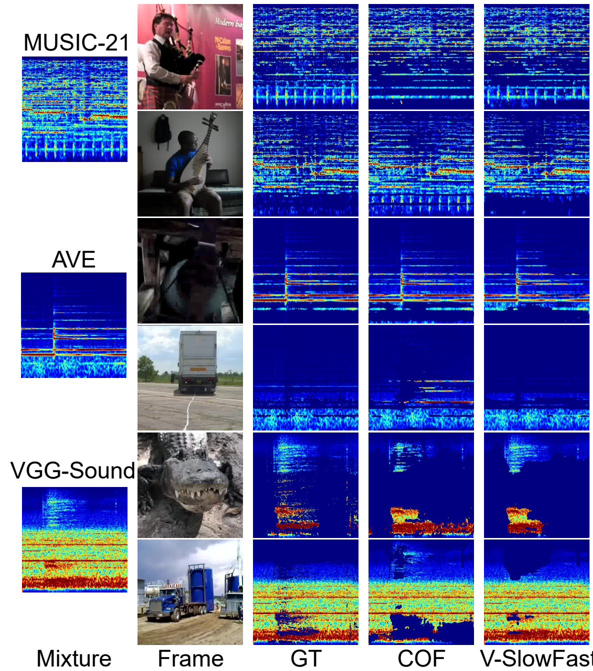

Figure 3. Visualization of the source separation results with the MUSIC-21, AVE, and VGG-Sound datasets.

which is multiplied with the original input mixture spectrogram to produce an estimate of the component audio corresponding to the visual input. More formally,

$$\begin{aligned} f_{Slow^{D,V}, n, \alpha_s} &= Slow^D\big(f_{Slow^{E,V}, n, \alpha_s}\big), \\ f_{Slow^{D,V}, n, 1} &= \phi^{inv}(f_{Slow^{D,V}, n, \alpha_s}, \alpha_s), \\ \hat{B}_{Slow,n,1} &= \sigma(f_{Slow^{D,V}, n, 1}), \\ \hat{X}_{Slow,n,1} &= \hat{B}_{Slow,n,1} \otimes X_{mix}, \end{aligned} \tag{5}$$

where the ϕ^{inv} represents the inverse ϕ operation. $f_{Slow^{D,V}, n, 1}$ indicates the separated sound features for the n-th source (n-th input frame) from the slow spectrogram network, which has the full temporal resolution. σ represents the sigmoid operation. $\otimes$ denotes the element-wise product. The output spectrogram $\hat{X}_{Slow,n,1}$ is formulated by element-wise multiplying the binary mask $\hat{B}_{Slow,n,1}$ with the original mixture spectrogram X_{mix}.

3.4. Fast Spectrogram Residual Network

The fast spectrogram residual network also uses an encoder-decoder structure. The coarsely separated spectrogram from the previous slow spectrogram network and the original mixture spectrogram are concatenated first then forwarded to the ϕ operation (with α_f, $\alpha_f < \alpha_s$). With the global attention from the visual embedding $e_{v,n}$, the fast spectrogram residual network produces a residual spectrum mask $f_{Fast^{D,V}, n, 1}$, which is added to the spectrum mask $f_{Slow^{D,V}, n, 1}$ (from the previous slow spectrogram network).

Models	MUSIC-21			AVE			VGG-Sound			Param (M)	GMACs
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR		
Sound of Pixels [64]	7.52	13.01	11.53	1.21	7.08	6.84	1.40	8.56	7.80	34.83	1.62
Minus-Plus [61]	9.15	15.38	12.11	1.96	7.95	8.08	1.93	9.30	8.25	58.35	5.13
Cascaded Opponent Filter (appearance) [65]	9.50	16.12	12.30	2.68	8.18	8.48	1.93	9.36	8.36	58.38	4.52
V-SlowFast (1)	**10.89**	**18.33**	**12.97**	**3.09**	**9.69**	8.45	**2.59**	**9.99**	8.69	32.22	2.17
V-SlowFast (2)	9.54	16.02	12.26	2.92	9.68	**10.08**	2.11	8.15	**13.35**	**15.07**	**0.84**

Table 1. Source separation performance using mixtures of two sources from the MUSIC-21, AVE and VGG-Sound datasets with the single-frame based models of Sound of Pixels [64], Minus-Plus [61], Cascaded Opponent Filter [65] and V-SlowFast. V-SlowFast (1) represents the model using V: MV2$^+$, Slow: DeepLabV3Plus and Fast: U-Net (9-layer). V-SlowFast (2) indicates the model of V: MV2$^+$, Slow: DeepLabV3Plus and Fast: DeepLabV3Plus. MV2$^+$ represents the vision network of *MV2 + AVGA + Contrast*. The V-SlowFast results of α_s=2, and α_f=1 are reported, and the overall best model results are **bolded**.

The final output is formed by multiplying the original mixture spectrogram with a binary mask obtained from the fast spectrogram residual network. More specifically,

$$\begin{aligned} f_{Fast^E,n,\alpha_f} &= Fast^E\big(\phi(cat[X_{mix}, \hat{X}_{Slow,n,1}], \alpha_f)\big), \\ f_{Fast^{E,V},n,\alpha_f} &= e_{v,n}^T \odot e_{v,n} \odot f_{Fast^E,n,\alpha_f}, \\ f_{Fast^{D,V},n,\alpha_f} &= Fast^D\big(f_{Fast^{E,V},n,\alpha_f}\big), \\ f_{Fast^{D,V},n,1} &= \phi^{inv}(f_{Fast^{D,V},n,\alpha_f}, \alpha_f), \\ \hat{B}_{Fast,n,1} &= \sigma(f_{Slow^{D,V},n,1} \oplus f_{Fast^{D,V},n,1}), \\ \hat{X}_{Fast,n,1} &= \hat{B}_{Fast,n,1} \odot X_{mix} \end{aligned} \tag{6}$$

3.5. Overall Learning Objective

The entire system is trained using a self-supervised setup with a large set of unlabelled videos. We formulate the visual sound separation learning objective to estimate the binary mask $\hat{B}_n$ to obtain the final output spectrogram (Eq. 3, 4, 5, and 6). The ground truth mask B_n is formed as,

$$B_n(f,t) = [\![X_n(f,t) \geq X_m(f,t)]\!] \tag{7}$$

where (f,t) represents the frequency-time coordinates in the sound spectrogram X. N is the number of sources in the mixture and $\forall m \in (1,\ldots,N)$. The **V-SlowFast** network is optimized by minimizing the binary cross entropy (BCE) loss between the estimated binary masks $\hat{B}_n$ and the ground-truth binary masks B_n,

$$\mathcal{L}_{sep} = \sum_{n=1}^{N} BCE(\hat{B}_{Slow,n,1}, B_n) + BCE(\hat{B}_{Fast,n,1}, B_n) \tag{8}$$

where $\hat{B}_{Slow,n,1}$ and $\hat{B}_{Fast,n,1}$ represent the predicted binary masks at slow spectrogram network and fast sectrogram residual network, respectively. Then the predicted mask is multiplied with the input full resolution mixture spectrogram to get a predicted sound spectrogram. Finally, we apply an inverse Short-time Fourier Transform (iSTFT) on the predicted spectrogram to reconstruct the waveform of separated sound. The overall learning objective is formed by combining the contrastive learning objective with the sound separation objective as $\mathcal{L} = \mathcal{L}_{contrast} + \mathcal{L}_{sep}$.

Models \ N	N=3			N=4		
	SDR	SIR	SAR	SDR	SIR	SAR
Sound of Pixels [64]	2.31	9.34	5.77	-0.22	6.99	3.80
Minus-Plus [61]	3.36	9.22	7.15	0.95	6.88	4.95
COF [65]	4.08	9.95	7.68	0.97	7.19	5.05
V-SlowFast (1)	**5.57**	**12.59**	**8.17**	**2.20**	**10.02**	**5.11**
V-SlowFast (2)	4.16	11.91	6.94	1.08	8.49	4.47

Table 2. Source separation performance using mixtures of three and four sources from the MUSIC-21 dataset using Sound of Pixels [64], Minus-Plus [61], COF [65] and V-SlowFast.

4. Experiments

In this section, we evaluate the visual sound separation performance of the proposed model.

4.1. Datasets and Implementation details

We train and evaluate the proposed methods using small- and large-scale datasets: MUSIC-21 [63], AVE [54] and VGG-Sound [12]. MUSIC-21 [63] contains 1365 videos from 21 instrumental categories. The AVE [54] dataset, a subset of AudioSet [27], contains 4143 10-second videos covering 28 audio-visual event categories. VGG-Sound [12] is a recently released large-scale dataset with over 200k video clips for 310 categories of general classes.

We follow the same setup as in [21, 67]. The datasets are split into disjoint train, val (not used for MUSIC-21), and test sets. The audio mixtures are obtained by adding the audio tracks from N videos (N depends on test setup). We apply temporal downsampling operation ϕ on spectrograms with a downsampling rate α (**Slow**: α_s and **Fast**: α_f). For instance, given a spectorgram with full temporal resolution of T, α_s=2 represents a slow spectrogram with temporal resolution of T/α_s. The sound separation performance is measured in terms of: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). For the measures of SDR and SIR, higher value indicates better performance (more details are presented in the supplementary material).

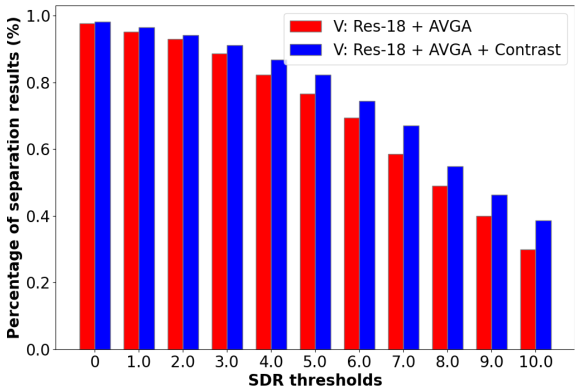

Figure 4. Percentage of separation results over SDR thresholds.

Models	SDR
Copy-Paste [67]	4.39
Sound of Pixels [64]	6.23
Minus-Plus [61]	7.11
V-SlowFast (1)	**8.64**
V-SlowFast (2)	8.51

Table 3. Separating sounds in MUSIC-21 from background noises.

4.2. Source Separation with V-SlowFast Network

Separating Two Sound Sources Table 1 summarizes the results in comparison with recent single frame methods Sound of Pixels [64], Minus-Plus [61] and Cascaded Opponent Filter (COF) [65] on MUSIC-21, AVE and VGG-Sound datasets using mixtures of two sound sources (N=2). We observe that our method consistently outperforms all baselines. Impressively, our system V-SlowFast (1) outperforms previous state-of-the-art multi-stage method [65] by 1.39dB on MUSIC-21, 0.41dB on AVE, and 0.66dB on VGG-Sound in terms of SDR while having substantially less parameters and small computational cost. Figure 3 illustrates qualitative examples and additional examples are provided in the supplementary material. V-SlowFast (2) can achieve similar result as multi-stage approach [65] and better performance than recursive model [61] while using 74.2% less parameters and 81.4% less operations. These quantitative and qualitative results suggest that our model successfully exploits the explicit slow and fast spectrogram separately to improve the sound separation quality and to substantially reduce the total model size and computation.

Separating More Sound Sources A more challenging task is to separate a sound mixture that contains more than two sources. To this end, we assess the approaches by separating mixtures of three and four sources using MUSIC-21 dataset. We report the separation performance of V-SlowFast and the baselines [64, 61, 65] in Table 2. V-SlowFast outperforms the baselines with a clear margin for separating mixtures of three and fours sources. Qualitative examples are provided in the supplementary material.

α	V	AVGA	Contrast	SDR	SIR	SAR	Param (M)	GMACs
	✓	✗	✗	8.06	14.79	10.82	36.03	2.04
α=1	✓	✓	✗	8.11	14.96	10.91	31.48	1.75
	✓	✓	✓	8.69	15.72	11.02	31.48	1.75
	✓	✗	✗	7.78	13.70	11.09	36.03	1.19
α=2	✓	✓	✗	8.02	14.46	11.11	31.48	1.04
	✓	✓	✓	8.61	14.90	11.35	31.48	1.04
	✓	✗	✗	7.03	12.30	11.05	36.03	0.77
α=4	✓	✓	✗	7.37	13.38	10.81	31.48	0.69
	✓	✓	✓	7.93	13.96	11.02	31.48	0.69
	✓	✗	✗	6.08	10.85	11.04	36.03	0.55
α=8	✓	✓	✗	6.43	11.11	11.17	31.48	0.52
	✓	✓	✓	6.93	11.83	11.43	31.48	0.52
	✓	✗	✗	4.88	8.85	11.33	36.03	0.45
α=16	✓	✓	✗	5.07	9.54	11.06	31.48	0.43
	✓	✓	✓	5.66	9.88	11.12	31.48	0.43

Table 4. Source separation performance using vision embeddings from vison network V of *Res-18*, *Res-18 + AVGA*, and *Res-18 + AVGA + Contrast*, and sound features from *U-Net (7-layer)* on mixtures of two sources from the MUSIC-21 dataset.

Separating Sound from Background Noises Due to the lack of ground truth for the source, assessing the performance in fully natural scenarios is difficult. However, we collect 100 natural background audios (retrieved from YouTube with keyword "background noise") to mix with available sources. Table 3 shows that our V-SlowFast models outperform all baselines on separating target sound from noisy mixture. "Copy-Paste" [67] uses input mixture as output. Note that the previous work of Cascaded Opponent Filter [65] requires the knowledge of all the presenting sources within the sound mixture to separate each sound source. Instead of relaying on the visual cues of other sources, our V-SlowFast model is proposed to efficiently separate the interested sound with only its associated visual information.

4.3. Ablation Study

Different Spectrogram Resolutions In Table 4, we report the visual sound separation performance on different spectrogram temporal resolutions ($\alpha \in \{1, 2, 4, 8, 16\}$). Similar as [64][1], the separation mask is obtained by a linear multiplication between the visual embedding (vision network: Res-18) and the sound features (sound spectrogram network: 7-layer U-Net). We observe that with a larger value of downsampling rate α (lower temporal resolution), the model converges earlier. The smaller temporal resolution the input spectrogram has, the lower evaluation scores of SDR and SIR the models obtain. More details are reported in the supplementary material.

[1] Differently, the spatial size of the U-Net encoder output is 16 times smaller than input spectrogram in our methods instead of 128 times in [64].

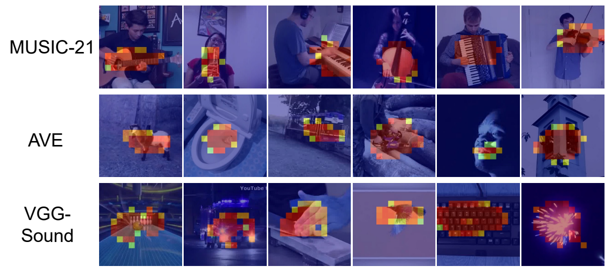

Figure 5. Visualization of the sounding sources with the MUSIC-21, AVE, and VGG-Sound datasets.

Audio-Visual Global Attention We assess the experiment results of using AVGA module together with vison network of Res-18 and sound spectrogram network of U-Net (7-layer) in Table 4. As we can see, adopting the AVGA module results in a smaller model and less computations while obtaining separation scores improvement for all the α, e.g. SDR gain of 0.34dB (α=4).

Visual Contrastive Learning We report the experiments result of using visual contrastive learning in Table 4 ($Contrast$). Note that the contrastive learning objective is only considered during training procedure. Thus, it does not bring extra operations for inference. As reported in Table 4, the visual contrastive learning improves the separation score by the gain of around 0.6dB in SDR for all the α.

For better visualizing the improvement the visual contrastive learning brings to *Res-18 + AVGA* model, we display the bar chart percentage of separation results over a wide range of SDR thresholds in Figure 4. The performance gap shows that the models using contrastive learning surpasss baseline with a large margin especially when the SDR threshold is $\geq$ 7.0. The contrastive learning allows the vision network to learn discriminative visual features and further improve the separation performance.

During inference, the sounding source location mask M_{loc} is yielded by an inner product between the visual feature and its own visual embedding (see Figure 2). Examples are reported in Figure 5 and in the supplementary material.

Architecture Variants of the Sound Network The architecture of the above mentioned sound spectrogram network is 7 up- and 7 down-convolutional layers U-Net, which is referred as U-Net (7-layer). In this section, we study how the model performs on different spectrogram resolutions when using less (5-layer) or more (9-layer) convolutional layers U-Net as the sound spectrogram network. Table 5 summarises the evaluation metrics, number of parameters and operations when using vision network of *Res-18 + AVGA + Contrast* and sound spectrogram network of U-Net (5-, 7-, 9-layer) with different α. We observe performance decrease when using shallower U-Net, and performance increase when using deeper U-Net for all the α. In addition, for smaller α (high temporal resolution), the performance increases a larger margin (e.g. 1.09dB in SDR of α=1) when switching from U-Net (7-layer) to deeper U-Net (9-layer) in comparison with counterparts of larger α (e.g. 0.58dB in SDR of α=16). Furthermore, using U-Net (9-layer) only increases total 2.22M parameters, and using U-Net (5-layer) outcomes a model with total 17.32M parameters.

$alpha$	U-Net	SDR	SIR	SAR	Param (M)	GMACs
α=1	5-layer	7.11	14.33	9.51	17.32	1.34
	7-layer	8.69	15.72	11.02	31.48	1.75
	9-layer	9.78	17.13	12.00	33.70	2.15
α=2	5-layer	6.96	13.62	9.99	17.32	0.84
	7-layer	8.61	14.90	11.35	31.48	1.04
	9-layer	9.55	16.09	12.10	33.70	1.25
α=4	5-layer	6.25	11.57	10.37	17.32	0.59
	7-layer	7.93	13.96	11.02	31.48	0.69
	9-layer	8.66	14.89	11.44	33.70	0.79
α=8	5-layer	5.71	10.54	10.83	17.32	0.47
	7-layer	6.93	11.83	11.43	31.48	0.52
	9-layer	7.60	12.75	11.82	33.70	0.57
α=16	5-layer	4.69	9.01	10.91	17.32	0.40
	7-layer	5.66	9.88	11.12	31.48	0.43
	9-layer	6.24	10.48	11.80	33.70	0.45

Table 5. Source separation performance using sound network of 5, 7, and 9 up- and down-convolutional U-Net layers with vision network of *Res-18 + AVGA + Contrast* on mixtures of two sources from the MUSIC-21 dataset.

Visually Guided SlowFast Sound Separation In this section, we examine the model performance when perceiving the slow and fast spectrograms separately.

V-SlowFast: We firstly separate sources using the V-SlowFast network **V**: Res-18$^+$, **Slow**: U-Net (7-layer), and **Fast**: U-Net (7-layer) in Table 6 (with $\alpha_s \in \{2,4,8,16\}$, $\alpha_f = 1$). The results with both of the slow and fast spectrograms clearly surpass the network with only single spectrogram model in Table 4, which proposes that treating the slow and fast spectrograms separately is important for the sound separation quality. Inspired by the observation from Table 5, that U-Net (5-layer) is a very light model and U-Net (9-layer) can have large performance gain without bring heavy parameters and operations, we design the architecture of the V-SlowFast by using **V**: Res-18$^+$, **Slow**: U-Net (5-layer), and **Fast**: U-Net (9-layer). As is shown in Table 6, with the compromise of around 0.2 $\sim$ 0.3 more GMACs, the method obtains performance gain (especially for the larger α_s) while having 11.95M less parameters for all the experiemnts. In comparison, we also examine how the "V-FastSlow" performs, where the fast spectropgram appears first and slow spectrogram occurs second. The V-FastSlow results in similar performance as the V-SlowFast in terms

V	Slow	Fast	SDR	SIR	SAR	Param (M)	GMACs
Res-18$^+$	α_s=2, U-Net (7-layer)	α_f=1, U-Net (7-layer)	10.43	17.88	12.62	51.69	2.45
	α_s=4, U-Net (7-layer)	α_f=1, U-Net (7-layer)	10.33	17.37	12.62	51.69	2.10
	α_s=8, U-Net (7-layer)	α_f=1, U-Net (7-layer)	9.69	16.84	12.12	51.69	1.93
	α_s=16, U-Net (7-layer)	α_f=1, U-Net (7-layer)	9.47	16.50	11.87	51.69	1.84
Res-18$^+$	α_s=2, U-Net (5-layer)	α_f=1, U-Net (9-layer)	10.55	18.10	12.54	39.74	2.65
	α_s=4, U-Net (5-layer)	α_f=1, U-Net (9-layer)	10.36	17.87	12.42	39.74	2.40
	α_s=8, U-Net (5-layer)	α_f=1, U-Net (9-layer)	9.94	17.27	12.30	39.74	2.28
	α_s=16, U-Net (5-layer)	α_f=1, U-Net (9-layer)	9.92	17.37	12.22	39.74	2.22
Res-18$^+$	α_s=4, U-Net (5-layer)	α_f=2, U-Net (9-layer)	9.90	17.06	12.14	39.74	1.50
	α_s=8, U-Net (5-layer)	α_f=2, U-Net (9-layer)	9.75	16.82	12.10	39.74	1.37
	α_s=16, U-Net (5-layer)	α_f=2, U-Net (9-layer)	9.57	16.71	11.86	39.74	1.31
	α_s=8, U-Net (5-layer)	α_f=4, U-Net (9-layer)	8.66	14.80	11.55	39.74	0.92
	α_s=16, U-Net (5-layer)	α_f=4, U-Net (9-layer)	8.54	14.80	11.36	39.74	0.86
Res-18$^+$	α_s=2, DeepLabV3Plus	α_f=1, U-Net (9-layer)	10.98	18.27	12.95	38.97	2.39
	α_s=4, DeepLabV3Plus	α_f=2, U-Net (9-layer)	10.38	17.29	12.59	38.97	1.37
MV2$^+$	α_s=2, DeepLabV3Plus	α_f=1, U-Net (9-layer)	10.89	18.33	12.97	32.22	2.17
	α_s=4, DeepLabV3Plus	α_f=2, U-Net (9-layer)	10.39	17.25	12.69	32.22	1.15
MV2$^+$	α_s=2, DeepLabV3Plus	α_f=1, DeepLabV3Plus	9.54	16.02	12.26	15.07	0.84
	α_s=4, DeepLabV3Plus	α_f=2, DeepLabV3Plus	8.64	14.89	11.54	15.07	0.48

Table 6. Source separation performance of the V-SlowFast framework using mixtures of two sources from the MUSIC-21 dataset. Res-18$^+$ represents the vision network of *Res-18 + AVGA + Contrast* and MV2$^+$ represents the vision network of *MV2 + AVGA + Contrast*.

of the evaluation metrics, number of parameters and operations (more details are presented in the supplementary material). Therefore, in this work, we discuss only on the case of V-SlowFast network.

Different combinations of α_s and α_f: We further study how the V-SlowFast performs on different combinations of α_s and α_f using **V**: Res-18$^+$, **Slow**: U-Net (5-layer), and **Fast**: U-Net (9-layer) in Table 6. We observe clear reduction in computations when using larger α_f (e.g. α_f=2, 4). The system has a slight performance drop with the increasing of the α_s, and the performance drops dramatically with the increasing of the α_f, which suggest that the performance of the fast spectrogram network determines the overall result.

Smaller architecture variants: In order to separate sound sources more efficiently, we explore the system with smaller architecture variants, e.g. MV2 [50]. We adapt MV2 as the vision network (4.52M parameters and 0.12 GMACs), DeepLabV3Plus [13] (MV2 as backbone) as the slow and fast specgrogram networks (5.27M parameters). The performance of the combinations between different model variants for the V-SlowFast framework are presented in Table 6. When using DeepLabV3Plus as the slow spectrogram network, and U-Net (9-layer) as the fast spectrogram residual network, the models with vision network of Res-18$^+$ and MV2$^+$ achieve similar results, e.g. Res-18$^+$: 10.98dB (α_s=2, α_f=1) and 10.38dB (α_s=4, α_f=2) of SDR in comparison with MV2$^+$: 10.89dB (α_s=2, α_f=1) and 10.39dB (α_s=4, α_f=2). Differently, the vision network of MV2$^+$ outcomes a model with 6.75M less parameters and 0.22 GMACs less operations. Thus, we refer the architecture of **V**: MV2$^+$, **Slow**: DeepLabV3Plus, and **Fast**: U-Net (9-layer) as **V-SlowFast** (1). Furthermore, when adopting the DeepLabV3Plus as the architecture of the fast spectrogram residual network, the model **V**: MV2$^+$, **Slow**: DeepLabV3Plus, and **Fast**: DeepLabV3Plus obtains close results as the recent single frame based state-of-the-art method COF [65] while only contains total 15.07M model parameters and 0.84 GMACs, which is denoted as **V-SlowFast** (2).

5. Conclusions

We proposed a new light yet efficient three-stream framework **V-SlowFast** that operates on visual image, slow spectrogram, and fast spectrogram. We introduced two contrastive objectives to encourage the network to learn discriminative visual features for separating sounds and localizing sounding sources. In addition, we proposed an audio-visual global attention module for audio and visual features fusion. Furthermore, we studied visually guided sound separation by treating the slow and fast spectrograms separately in terms of different temporal resolutions and model variants. The proposed **V-SlowFast** models show excellent performance on small- and large-scale datasets: MUSIC-21, AVE, and VGG-Sound.

Acknowledgement This work is supported by the Academy of Finland (projects 327910 & 324346).

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020.

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016.

[6] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[8] 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 (CVPR)*, July 2017.

[9] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.

[10] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020.

[11] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[12] 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.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020.

[16] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shunichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[19] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[20] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020.

[21] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[22] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020.

[23] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018.

[24] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019.

[25] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3879–3888, 2019.

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

[27] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[28] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pages 472–478, 1996.

[29] Elana Zion Golumbic, Gregory B Cogan, Charles E Schroeder, and David Poeppel. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, 33(4):1417–1426, 2013.

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020.

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Di Hu, Dong Wang, Xuelong Li, Feiping Nie, and Qi Wang. Listen to the image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7972–7981, 2019.

[34] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 425–442. Springer, 2020.

[35] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[36] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021.

[37] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

[38] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019.

[39] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021.

[40] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global and local audio-visual representations. *arXiv preprint arXiv:2104.05418*, 2021.

[41] David McQuillan. The feasibility of ludgate's analytical machine.

[42] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. *arXiv preprint arXiv:2011.01819*, 2020.

[43] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *arXiv preprint arXiv:1809.02587*, 2018.

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[45] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[46] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020.

[47] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[49] Sam T Roweis. One microphone source separation. In *Advances in neural information processing systems*, pages 793–799, 2001.

[50] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[51] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[52] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[53] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[54] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[55] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[56] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021.

[57] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[58] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021.

[59] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[60] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[61] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[62] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020.

[63] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[64] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[65] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation using cascaded opponent filter network. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[66] Lingyu Zhu and Esa Rahtu. Leveraging category information for single-frame visual sound source separation. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2021.

[67] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. *arXiv preprint arXiv:2104.08506*, 2021.