Visually Guided Audio Source Separation with Meta Consistency Learning

Md Amirul Islam[1] Seyed Shahabeddin Nabavi[1] Irina Kezele[1]
Yang Wang[2] Yuanhao Yu[1] Jin Tang[1]
[1]Huawei Noah's Ark Lab, [2]Concordia University

Abstract

In this paper, we tackle the problem of visually guided audio source separation in the context of both known and unknown objects (e.g., musical instruments). Recent successful end-to-end deep learning approaches adopt a single network with fixed parameters to generalize across unseen test videos. However, it can be challenging to generalize in cases where the distribution shift between training and test videos is higher as they fail to utilize internal information of unknown test videos. Based on this observation, we introduce a meta-consistency driven test time adaptation scheme that enables the pretrained model to quickly adapt to known and unknown test music videos in order to bring substantial improvements. In particular, we design a self-supervised audio-visual consistency objective as an auxiliary task that learns the synchronization between audio and its corresponding visual embedding. Concretely, we apply a meta-consistency training scheme to further optimize the pretrained model for effective and faster test time adaptation. We obtain substantial performance gains with only a smaller number of gradient updates and without any additional parameters for the task of audio source separation. Extensive experimental results across datasets demonstrate the effectiveness of our proposed method.

1. Introduction

Recent advancements of convolutional neural networks (CNNs) for audio-visual source separation [16, 17, 43, 50, 55, 56, 62] have lead to significant boost in performance. However, separating sounds in the wild is still a challenging problem due to insufficient separation cues in a mixture of sounds. Existing approaches use visual information (e.g., lip movements in speech separation [17], musical instrument type in audio separation [4, 16, 23, 43, 50, 54–56, 62], etc.) to learn an audio-visual representation to promote the separation process. Furthermore, incorporating other modalities (e.g., pose [43], motion [36, 55, 62], keypoints [14]) is shown to improve the separation performance by learning a multi-modal representation. In this work we follow the setup from similar prior works [16, 23, 43, 50, 54–56, 62], where all the sounds in the video can be related to on-screen objects. Despite their success, these methods are limited to the scenario where both train and test sets contain similar objects (e.g., musical instruments).

While music videos in the wild can include various musical instruments and different music styles and musical expressiveness, currently available datasets cover only a limited set of music video examples and only a part of the comprehensive list of the instrument categories. Working with such datasets, prior works have demonstrated a limited generalization capability to unknown music videos and instrument categories at test time, and a failure to overcome the distribution gap between the training and testing data.

The main motivation of our work is to propose an adaptation framework to improve generalization to known and unknown test videos, and in an extension, to specifically promote generalization to new sound categories (e.g., new music instruments). To achieve this, at test time we exploit the internal information of each individual test sample.

Adapting a model on unknown test samples termed as "online matching" or "test time adaptation" has shown to be effective in various tasks (e.g., image recognition [5, 6, 34], image deblurring [7], video [3]). However, most of these methods adopt test time adaptation in a naive way [25, 48, 58] which requires considerable inference time or increased amount of parameters. In addition, naive test time adaptation can drive the model to catastrophic forgetting [30]. Meta-learning techniques, such as model agnostic meta-learning (MAML) [12] have been introduced to remedy such limitations in test time adaptation. However, the idea of meta-learning driven test time adaptation has not been explored in the context of audio visual learning.

Motivated by existing works [6–8, 37], we combine the idea of meta-learning [12] with self-supervised auxiliary learning. To this end, we introduce a meta-auxiliary based pipeline for visually guided audio source separation which can quickly adapt to an unseen music video at test time, to substantially improve the separation performance. In particular, we first design a self-supervised cross-modal consistency objective [17, 26, 32, 58] as an auxiliary task that learns

the synchronization between audio and its corresponding visual embedding. Note that both the primary task (i.e, source separation) and the auxiliary task (i.e., audio-visual consistency learning) share majority of the parameters. The model parameters are learned in a meta-learning fashion, so that the updated model performs well on the primary task after adaptation using the consistency objective. However, as mentioned earlier, naively updating the model parameters via the consistency objective can lead to catastrophic forgetting [30]. Therefore, we take one step further towards the generalization of sound separation methods on a more advanced setting, where the musical instruments in training and testing are not the same in a general case. The contributions of our paper are as follows:

- We propose a novel framework for visually guided audio source separation in the context of both known and unknown musical instruments.
- We introduce an effective meta-consistency driven test time adaptation scheme that enables the pretrained model to quickly adapt to unknown music videos. To the best of our knowledge, the proposed approach is the first work on meta-auxiliary driven test-time adaptation in the context of audio-visual learning.
- We provide experimental evidence that our method improves generalization to both known and unknown musical categories. We further demonstrate consistent improvements on existing state-of-the-art methods.

2. Related Work

Audio-visual Sound Separation. Recent deep learning based approaches separate visually indicated sounds for various sources including speech [1, 2, 9, 11, 13, 17, 35, 41], objects [15], musical instruments [4, 14, 16, 23, 43, 50, 54–56, 59–62], and universal purposes [15, 45]. Zhao *et al*. [56] introduced PixelPlayer, a framework to learn object sounds and their location in the scene for sound source separation. Gao *et al*. [16] introduced a novel co-separation objective to associate consistent sounds to the objects of the same category across all training samples. Tian *et al*. [50] proposed sounding object visual ground network along with a co-learning paradigm to determine if the object is audible to further separate its source. Zhu *et al*. [61] adapted the classical slowfast networks to propose a three-stream slowfast network along with a contrastive objective. Another line of works focused on incorporating additional input modalities (e.g., pose [43], motion [36, 55, 62], keypoints [14]) to improve the source separation performance by learning a multi-modal representation. In particular, TriBERT [43] devised a multi-modal transformer to utilize pose information along with weak category supervision on pose and visual embedding for fully supervised audio visual source separation. Our base network for visually guided audio separation is inspired by TriBERT [43] and VilBert [28]. However, unlike TriBERT, our method does not necessitate delineation of visual ROIs for tokenization by weakly supervised segmentation, leading to a simpler training scheme, yet achieving a superior performance. In addition, these approaches adopted a single network with fixed parameters to generalize across unseen test videos, while we introduce a self-supervised audio-visual consistency objective as an auxiliary task that learns synchronization between audio and its corresponding visual embedding. Similar to our work, Zhou *et al*. [58] proposed SeCo framework to separate unknown musical instruments by exploiting the consistency constraints in online matching strategy, which can bring stable enhancements with no cost of extra parameters. However, SeCo [58] is neither designed nor trained for adaptation in a low-data regime (one sample is the limit).

Auxiliary and Meta Learning. The goal of auxiliary learning is to enhance the generalization of the primary task [27]. The auxiliary task is employed for various purposes including depth completion [29], super resolution [38], and deblurring [7]. In addition, meta-learning enables fast test time adaptation via a few training examples. The idea of combining auxiliary learning with meta learning [12] is already explored in existing works [6–8, 37]; however, it is not explored in the context of audio visual learning. To the best of our knowledge, we are the first to apply meta-consistency training scheme to further optimize the pretrained model for effective test time adaptation.

3. Audio Visual Source Separation Network

We tackle the visual sound separation task where the goal is to separate sounds from a mixed audio signal by incorporating visual information. Following [17, 21, 50, 56, 62], we adopt the "mix and separate" technique to train the model. Given two video clips $\{V_1, V_2\}$ with associated audio signals $\{A_{V_1}, A_{V_2}\}$, we first mix the audio signals to generate a synthetic mixed signal, $A_{V_{mix}} = A_{V_1} + A_{V_2}$. Following common practice [16, 17, 43, 50, 55, 56, 62], we apply short time Fourier transform (STFT) [19] on the raw mixed signal $A_{V_{mix}}$ to generate a log spectrogram, S_m, for the ease of training. The overall architecture of our network is illustrated in Fig. 1. The visual guidance network (Sec. 3.2) guides the encoded feature of the audio separation network (Sec. 3.1) which outputs separated audio spectrogram masks. We multiply the separated audio masks by the mixed spectrogram, and apply inverse STFT to generate the clean separated audio signals. To make our model adaptive to specific test samples, we customize it to utilize additional internal information of each test sample separately, which is available at test time. Towards this goal, we first design a self-supervised cross-modal consistency objective

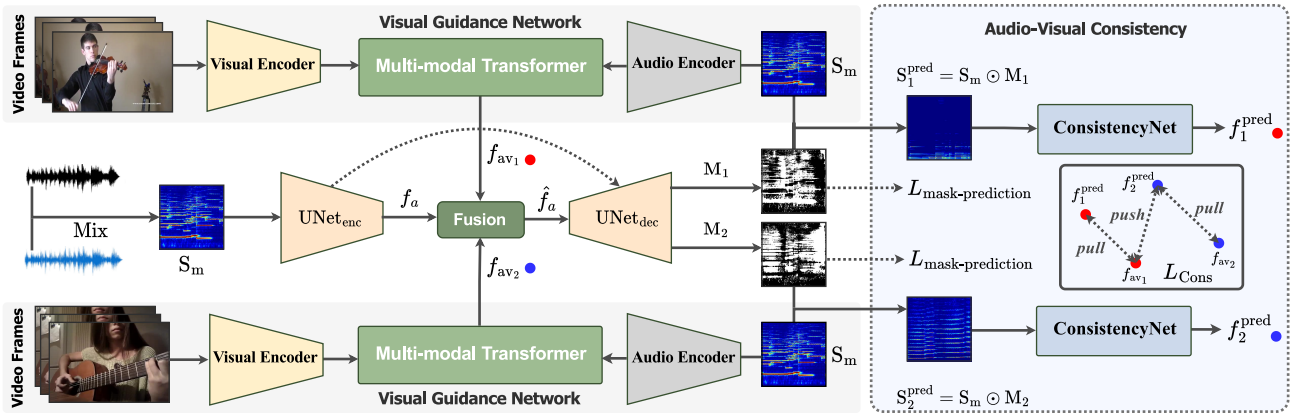

Figure 1. Illustration of our architecture for visually guided sound separation. The visual guidance network (Sec. 3.2) extracts multi-modal representation, f_{av}, from the video frames and mixed spectrogram, S_m. The mixed spectrogram is also fed to the audio separation network (Sec. 3.1) which generates the encoded audio feature, f_a. Then, the multi-modal representation, f_{av}, is combined with the audio feature, f_a, resulting in a guided audio feature map, $f_{\hat{a}}$. The decoder of the audio separation network takes $f_{\hat{a}}$ as input and outputs separated audio spectrogram masks, $\{\mathbf{M}_{V_1}, \mathbf{M}_{V_2}\}$. We then design a self-supervised cross-modal consistency objective (Sec. 3.3) as an auxiliary task that learns synchronization between the audio and its corresponding visual embedding.

(Sec. 3.3) as an auxiliary task to learn audio-visual synchronization. Finally, we apply meta-consistency training scheme (Sec. 3.4) to further optimize the pretrained model for effective test time adaptation.

3.1. Audio Separation Network

Following existing works [16, 17, 43, 55, 56, 58], we use an attention U-Net [44] style encoder-decoder network with skip connection to generate separated audio spectrograms. Note, the U-Net contains seven convolutions and seven deconvolutions layers. The encoder of attention U-Net takes a mixed audio spectrogram $S_m \in \mathbb{R}^{1\times256\times256}$ as input and extracts an audio feature map $f_a \in \mathbb{R}^{1024\times16\times16}$. The encoded representation, f_a, is combined with the multi-modal representation f_{av} (obtained from the visual guidance network described in Sec. 3.2) with a self-attention based fusion technique used in [14, 43]. Note that before the fusion, we adjust the dimension of audio and multi-modal features. The fused feature $f_{\hat{a}}$ is fed into the decoder of the attention U-Net which predicts the final magnitude of the spectrogram masks, $\{\mathbf{M}_1, \mathbf{M}_2\}$. Finally, we activate the predicted spectrogram masks via the sigmoid function to obtain the predicted separation masks $\{\mathbf{M}^{S_1}, \mathbf{M}^{S_2}\}$. For the sound separation task, since our goal is to separate spectrogram masks for each individual object, we apply separation loss, L_{mask}, between the predicted separated masks and the binary ground-truth masks. L_{mask} uses a per-pixel sigmoid cross entropy objective. Following prior works [43, 50, 56, 62] the binary ground-truth mask of each video is calculated by observing whether the target sound is the dominant component in the mixture sound. This loss provides the main supervision to enforce the separation of the clean audio.

3.2. Visual Guidance Network

Inspired by the success of multi-modal transformer in various tasks [28, 31, 41, 43, 49], we build our visual guidance network on top of ViLBERT [28] and TriBERT [43]. ViLBERT is a two stream architecture which jointly learns from image and text while TriBERT extends ViLBERT's architecture to three stream (vision, audio, and pose) to learn a human-centric audio-visual representation. In contrast, we build a two-stream visual guidance network to learn an audio-visual representation. Unlike visual guidance networks in existing works [16, 17, 50, 55, 56, 62] which only use visual cues, our guidance network takes both video frames and mixed spectrogram as input, and outputs a joint audio-visual representation. The joint representation is used to guide the audio separation network. Similar to [28, 41, 43], we use a bi-directional transformer encoder [51] as the backbone of the guidance network. We first generate the visual *tokens* by directly feeding video frames to a CNN architecture. We then apply a tiny VGG network [46] on mixed audio spectrogram to generate the audio *tokens*. The two sets of tokens are fed to the multi-modal transformer encoder, which refines them using bi-modal co-attention, to output a multi-modal representation.

Visual Representations. TriBERT uses an end-to-end segmentation network which outputs detected object features to feed into multi-modal transformer. In contrast, we directly use input frames for each video separately to extract global semantic representation rather than using detected bounding box features [16, 43, 50]. We use the 2D ResNet-50 architecture [20] as the visual analysis network which takes a video as input, and outputs 1024 dimensional feature vector after the last spatial average pooling layer. We then reshape the feature vector along the temporal dimension (i.e, number of

video frames) and the resultant visual embedding is fed to the multi-modal transformer as visual *token*.

Audio Representations. The mixed audio spectrogram, $S_m \in \mathbb{R}^{1\times256\times256}$ (used in Sec. 3.1) is fed to a tiny VGG-like [46] architecture which outputs the high-level global audio embedding. The audio embedding is repeated to generate audio sequences which are used as *tokens* for audio stream of the multi-modal transformer.

Bi-modal Co-attention. Following [28, 41, 43], we use bi-modal co-attention layer in the transformer encoder to learn effective representations. While TriBERT [43] extends the ViLBERT's co-attention layer to take intermediate representation of three different modalities, we extend it to take intermediate vision and audio representation as input. We keep the rest of the transformer encoder architecture similar to ViLBERT [28]. The resultant audio-visual representation is used to guide the encoded features from the audio separation network. Note that our guidance network does not use any audio level category information or other modality (i.e., pose) as used in TriBERT [43].

3.3. Cross-Modal Consistency Network

In addition to the audio visual source separation network, a properly chosen self-supervised *auxiliary* task can complement the *primary* separation task in a way that can be used to adapt the network on test samples. Unlike the original setting (i.e., separating the known musical instruments), we are also interested in exploring a more challenging scenario to separate the unknown musical categories by achieving stronger adaptation ability. Motivated by existing works [24, 26, 32, 58], we introduce an audio-visual consistency analysis network which learns the synchronization of video and corresponding separated audio. The consistency network is likely to capture the audio-visual correlation when adapted to new samples leading to better audio separation result. Note that the auxiliary audio-visual consistency task is self-supervised.

To learn audio-visual synchronization, we use inter-modal consistency [17, 26, 32, 58] based on the predicted audio masks from the separation network. We multiply the predicted audio spectrograms $\{\mathbf{M}_1, \mathbf{M}_2\}$ by the mixed spectrogram, $\mathbf{S}_m$, to obtain the separated audio spectrograms, $\{\mathbf{S}_1^{\text{pred}}, \mathbf{S}_2^{\text{pred}}\}$. For consistency computation, we further use a ResNet18 [20] to encode the two predicted audio spectrograms into a lower dimensional embedding space suitable for direct comparison with the visual embeddings obtained by the visual guidance network. The consistency network takes the separated spectrograms, $\{\mathbf{S}_1^{\text{pred}}, \mathbf{S}_2^{\text{pred}}\}$, as input and outputs 256 dimensional consistency embedding, $\{\mathbf{f}_1^{\text{pred}}, \mathbf{f}_2^{\text{pred}}\}$ for each spectrogram separately. Similar to [17, 26, 58], the audio-visual associations in videos are learned in a straight-forward efficient way, where the training objective is to minimize the distance of the positive pairs while maximizing the distance for negative pairs. We consider the synchronized audio-visual samples (i.e, the separated audio embedding and their corresponding visual embedding) as a positive pair $\{\mathbf{f}_i^{\text{pred}}, \mathbf{f}_i^{\text{v}}\}$ where $i \in \{1, 2\}$. In contrast, we obtain the negative pairs by cross pairing the audio and visual embedding, $\{\mathbf{f}_i^{\text{pred}}, \mathbf{f}_j^{\text{v}}\}$ where $i \neq j$ & $(i, j) \in \{1, 2\}$. We normalize all the embeddings before consistency computation using sigmoid function. The overall audio-visual consistency loss can be defined by the following:

$$\begin{aligned} L_{\text{Cons}} = L_2(\mathbf{f}_1^{\text{pred}}, \mathbf{f}_1^{\text{v}}) + L_2(\mathbf{f}_2^{\text{pred}}, \mathbf{f}_2^{\text{v}}) \\ -L_2(\mathbf{f}_1^{\text{pred}}, \mathbf{f}_2^{\text{v}}) - L_2(\mathbf{f}_2^{\text{pred}}, \mathbf{f}_1^{\text{v}}) \end{aligned} \tag{1}$$

This loss forces the overall network to learn cross-modal visual audio embeddings such that the distance between the embedding of the separated music and the visual embedding for the corresponding musical instrument should be smaller than that between the separated audio embedding and the visual embedding for the other musical instrument.

Note that the separation results at the beginning of training are not sufficiently accurate to learn audio-visual association, in fact they can confuse the network to identify positive and negative pairs reliably. To address this limitation, following [26, 58], we incorporate ground-truth audio features to help the association learning process in the beginning of the optimization process. More specifically, we pass the ground-truth audio masks to the consistency network and generate embeddings $\{\mathbf{f}_1^{\text{GT}}, \mathbf{f}_2^{\text{GT}}\}$ to include an additional loss term, $\lambda = \delta(L_2(\mathbf{f}_1^{\text{GT}}, \mathbf{f}_1^{\text{v}}) + L_2(\mathbf{f}_2^{\text{GT}}, \mathbf{f}_2^{\text{v}}))$ for regularization in Eq. 1. Note that the regularizer λ is only used at the beginning of the learning process to help "warm-start" the learning. It is excluded during meta-consistency training and test time adaptation in Sec. 3.4. The weight δ decays fast over the course of training (see S2 in supplement). The overall loss function for training is as follows:

$$L = L_{\text{mask}} + \gamma * (L_{\text{Cons}} + \lambda) \tag{2}$$

where γ is the weight for the consistency loss.

3.4. Meta-Consistency Learning for AVSS

Existing works [3, 5, 52, 58] utilize online matching strategy also termed as 'test-time adaptation' which adapts a learned model to unknown samples during inference. This is achieved by fine-tuning the model parameters for each test sample based on the error signals from a self-supervised *auxiliary* loss. However, there exist works [6–8, 37, 53, 57] which pointed out that naively applying test-time adaptation as in [48, 58] can lead to catastrophic forgetting as the parameters updated via self-supervised loss are biased towards improving the *auxiliary* self-supervised task rather than the *primary* task. To address this limitation, some works [6–8,

37,53,57] introduced a learning framework which integrates meta learning with *auxiliary* self-supervised learning. Motivated by these works [6–8, 37, 53, 57], we introduce meta-consistency training framework for audio-visual source separation, with the goal of further improving the separation results and adapting to known/unknown samples.

The overall meta-consistency learning pipeline is presented in Algorithm 1. We first initialize the parameters from the pre-trained audio-visual separation model which is already capable of separating audios. During meta-consistency learning, we enforce the constraint that the parameter update via the cross-modal consistency loss (Eq. 1) should improve the audio separation task. We now describe the flow of our algorithm in more detail. We decompose the parameters of our entire model as $\theta = \{\theta^S, \theta^P, \theta^{\text{Cons}}\}$, where θ^S denotes the shared weights, θ^P and θ^{Cons} are the weights for the *primary* source separation branch and the *auxiliary* audio-visual consistency branch, respectively. Note that θ^P is also required for our *auxiliary* task, since the *auxiliary* consistency task uses the output from the *primary* separation task. We denote the gradient update iterations for each sample as the inner loop and the meta-update iterations as the outer loop. During the inner loop training, given an audio-visual pair and the parameters of pretrained model θ, we perform a small number of gradient updates on the input pair using the consistency loss:

$$\hat{\theta}_n = \theta - \alpha \nabla_\theta L_{\text{Cons}}(\mathbf{f}_n^{\text{pred}}, \mathbf{f}_n^{\text{v}}; \theta), \tag{3}$$

where α is the adaptation learning rate. $\mathbf{f}_n^{\text{pred}}, \mathbf{f}_n^{\text{v}}$ refer to audio and visual embeddings, respectively. Here, $\hat{\theta}_n$ involves all the model parameters, $\{\hat{\theta}_n^S, \hat{\theta}_n^P, \hat{\theta}_n^{\text{Cons}}\}$. Our goal is to force the updated $\{\hat{\theta}_n^S, \hat{\theta}_n^P\}$ to enhance the audio separation task by minimizing the separation loss, L_{mask}. Following [7, 8, 37, 57], we define the meta-objective as:

$$\min_{\theta^S, \theta^P} \sum_{n=1}^{N} L_{\text{mask}}(\mathbf{M}_n^{\text{S}}, \mathbf{M}_n^{\text{gt}}; \hat{\theta}_n^S, \hat{\theta}_n^P), \tag{4}$$

where L_{mask} is a function of $\hat{\theta}_n$ but the optimization is over θ. $\mathbf{M}_n^{\text{S}}, \mathbf{M}_n^{\text{gt}}$ refer to the predicted and the ground-truth audio masks, respectively. The meta-objective in Eq. 4 can be minimized as following:

$$\theta \leftarrow \theta - \beta \sum_{n=1}^{N} \nabla_\theta L_{\text{mask}}(\mathbf{M}_n^{\text{S}}, \mathbf{M}_n^{\text{gt}}; \hat{\theta}_n^S, \hat{\theta}_n^P), \tag{5}$$

where β is the meta learning rate and following existing practice [6–8, 37, 53, 57], we use a mini-batch in Eq. 5. Note that we only update consistency network parameters, θ^{Cons}, in the inner loop while update audio separation network parameters, θ^S and θ^P, in the outer loop.

Algorithm 1: Meta-Consistency Learning

Input: Consistency pretrained model parameters
Output: Meta-consistency learned parameters, θ

Initialize the model with pre-trained parameters: $\theta = \{\theta^S, \theta^P, \theta^{\text{Cons}}\}$
while *not converged* **do**
  Sample a batch, N of audio-visual pairs $\{\text{I}^{\text{a}}, \text{I}^{\text{v}}\}$
  foreach *pair* $(I_n^a, I_n^v) \in N$ **do**
    while *iter* $\leq$ *number of updates,* k **do**
      Evaluate $\nabla_\theta L_{\text{Cons}}$ using Eq. 1
      Compute adapted parameters with GD: $\hat{\theta}_n = \theta - \alpha \nabla_\theta L_{\text{Cons}}(\mathbf{f}_n^{\text{pred}}, \mathbf{f}_n^{\text{v}}; \theta)$
      Update: $\theta^{\text{Cons}} \leftarrow \theta^{\text{Cons}} - \alpha \nabla_\theta L_{\text{Cons}}(\mathbf{f}_n^{\text{pred}}, \mathbf{f}_n^{\text{v}}; \theta^{\text{Cons}})$
  Evaluate the audio separation task and update: $\theta \leftarrow \theta - \beta \sum_{n=1}^{N} \nabla_\theta L_{\text{mask}}(\mathbf{M}_n^{\text{S}}, \mathbf{M}_n^{\text{gt}}; \hat{\theta}_n^S, \hat{\theta}_n^P)$

Meta auxiliary testing. During meta-testing, given an audio visual pair, we obtain the adapted parameter $\hat{\theta}$ by simply applying Eq. 3. The final separation masks are obtained from the adapted parameters, $\hat{\theta}$. The model parameters are switched back to the original meta-trained state before evaluating the next pair.

4. Experiment

4.1. Implementation Details

We implement our pipeline using the PyTorch framework [39]. Following previous works [43, 50, 56], we consider three consecutive RGB frames with a spatial resolution of 224×224 as an input sequence to our visual stream of the multi-modal transformer. We set the video frame rate to 1fps and randomly sample 3 consecutive frames from 6s video clip. Similar to existing works [43, 50, 56], we subsample 6s of audio signals at 11KHz from the same clip, to reduce the computational cost. We use STFT with a hop length of 256 and Hann window size of 1024 for the spectrogram generation process. We use an ImageNet [10] pretrained ResNet50 architecture to extract visual features. We use the BERT Adam optimizer with an initial learning rate of $1e^{-5}$ and batch size of 12 to train our base model on two Tesla V100 GPUs for 100 epochs. The weight γ in Eq. 2 is set to 0.01. During the meta-consistency training, we set the learning rate to $1e^{-9}$ and $1e^{-10}$ for the outer loop update and the inner loop updates, respectively. During meta-testing, we set the learning rate to $2e^{-5}$. We follow the line of works [50, 56, 62] which run inference on pairs of test videos. We report numbers for the following variants that are described in what follows: **AVSS:** This is our baseline audio visual source separation net-

work which consists of visual guidance network (Sec. 3.2) and audio separation network (Sec. 3.1). **AVSS+CMC:** This network applies the cross-modal consistency module (Sec. 3.3) with AVSS. **AVSS+CMC+Meta:** This network applies meta-consistency training scheme (Sec. 3.4) to AVSS+CMC. **AVSS+CMC+Naive TTA:** This variant applies test time adaptation on the trained AVSS+CMC (i.e., test time adaptation without meta-consistency learning) to demonstrate that TTA performs well with a meta-trained model. **AVSS+CMC+Meta TTA:** This is our final method which applies meta-consistency driven test time adaptation on the trained AVSS+CMC+Meta.

4.2. Dataset and Evaluation Metrics

Dataset. We quantitatively evaluate our model on MIT MUSIC [56] and MUSIC-21 [55] datasets. The MUSIC dataset consists of 11 classes of musical instrument and composed of untrimmed videos crawled from the YouTube. Following CCoL [50], we gather 510 online available musical solo videos from the MUSIC dataset and split it into training/validation/test sets, which have 420/45/25 videos from different categories, respectively. The MUSIC-21 dataset consists of 21 classes and contains 1365 untrimmed videos. For a fair comparison, we use the split provided by TriBERT [43] to report results on the MUSIC-21 dataset. Note that we include the evaluation performance of our method and the baselines when we use only single-source videos (solos) or multi-source videos (solos+duets).

Evaluation Metrics. We use the widely used mir eval library [40] to quantify the performance under three standard metrics: Signal-to-distortion ratio (SDR), Signal-to-interference ratio (SIR), and Signal-to-artifact ratio (SAR).

4.3. Quantitative Results on MUSIC

We compare our approach with recent state-of-the-art methods under two different MUSIC test splits for audio-visual source separation task. Table 1 and Table 2 summarize the results for separating two sound sources on the split provided by [50] and [16], respectively. As we can see from Table 1, our approach outperforms (11.77dB vs 7.27dB in SDR) the compared methods by large margins in terms of SDR and SIR. Similarly, our proposed model outperforms its closest competitor (Table 2) by substantial margins of around 0.5dB SDR and 1.3dB SIR on the MUSIC test set. Figure 2 illustrates the qualitative comparison results. It is clear that our approach, both quantitatively and qualitatively, outperforms the baselines in sound separation.

To further demonstrate the superiority of our approach, we follow [43] to report the results on MUSIC test set under the setting when we first train on MUSIC-21 dataset and then apply meta-consistency learning scheme (Alg. 1) on MUSIC dataset. For this setting, we follow the split provided by Co-Separation [16] for a fair comparison. Table 3 summarizes the comparison results. Our method consistently outperforms the baselines in separation accuracy for most cases with a large margin. Note that TriBERT [43] requires human pose keypoints as an input, while our approach does not require this additional information. The main limitation of test time adaptation is that it comes at the cost of longer inference time; however, in most cases, we demonstrated substantial improvements with only one or two gradient updates. This slight increase in inference time is balanced well with significant generalization boost, particularly in the case of unknown test categories.

Method	SDR ↑	SIR↑
RPCA [22]⋆	-0.48	3.13
Sound of Pixels [56]⋆	3.42	4.98
Co-Separation [16]⋆	2.04	6.21
CCoL [50]⋆	7.27	12.77
Ours	**11.77**	**19.36**

Table 1. The separation performance comparison on MUSIC test split provided by [50]. The results indicated with ⋆ are obtained from [50]. Our method achieves SOTA in SDR and SIR metrics.

Method	SDR ↑	SIR↑	SAR↑
Sound of Pixels [56]⋆	6.1	10.9	10.6
Minus-Plus Net [54]⋆	7.0	14.4	10.2
Sound of Motion [55]⋆	8.2	14.6	13.2
Co-Separation [16]⋆	7.4	13.8	10.6
Music Gesture [14]⋆	10.1	15.7	12.9
AVSGS [4]⋆	11.4	17.3	**13.5**
Ours	**11.9**	**18.6**	**13.5**

Table 2. The audio separation performance comparison on MUSIC test split provided by [16]. The results indicated with ⋆ are obtained from [4]. Our method outperforms all the existing approaches in terms of SDR and SIR metrics and levels with the best previous approach for SAR.

Method	SDR ↑	SIR↑	SAR↑
NMF-MFCC [47]⋆	0.92	5.68	6.84
AV-Mix-and-Separate [16]⋆	3.23	7.01	9.14
Sound of Pixels [56]⋆	7.26	12.25	11.11
Co-Separation [16]⋆	7.64	13.8	11.3
Mask Co-efficient [42]⋆	9.29	15.09	12.43
TriBert [43]⋆	12.34	18.76	**14.37**
Ours	**12.81**	**19.56**	14.16

Table 3. The performance comparison on MUSIC test split [16] when the models are pretrained on MUSIC-21 dataset. The results indicated with ⋆ are obtained from [43]. Our method outperforms all the existing approaches in terms of SDR and SIR metrics, and is somewhat outperformed only by TriBERT in the SAR metric.

4.4. Adaptation Results on MUSIC-21

To show model compatibility to separate sounds of unknown musical instrument, we evaluate our framework on the MUSIC-21 dataset [55]. More specifically, we use the

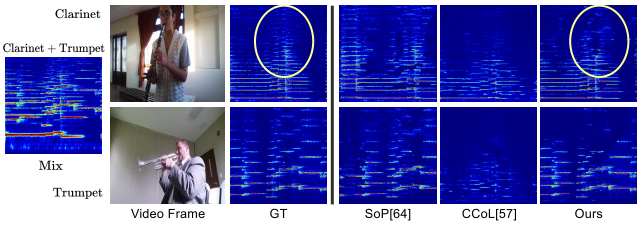

Figure 2. Qualitative sound separation results on the MUSIC test set. Here, we show a comparison with SoP [56] and CCoL [50].

Method	SDR ↑	SIR↑	SAR↑
AVSS	3.38	9.28	8.45
AVSS + CMC	3.98	9.85	8.91
AVSS + CMC + Naive TTA	4.05	10.0	8.77
AVSS + CMC + Meta TTA	**4.43**	**10.26**	**9.05**

Table 4. Adaptation results on the MUSIC-21 **multi-source** set.

Method	SDR ↑	SIR↑	SAR↑
AVSS	3.87	9.98	8.73
AVSS + CMC	4.70	10.93	9.19
AVSS + CMC + Naive TTA	4.75	10.96	9.20
AVSS + CMC + Meta TTA	**4.94**	**11.07**	**9.22**

Table 5. Adaptation results on the MUSIC-21 **single-source** set.

meta trained model from MUSIC dataset and adapt it on each sample of the MUSIC-21 test set. For each test pair, we optimize the meta-trained model parameters using the cross-modal consistency loss for several inner loop iterations. Unlike SeCo [58] which showed the adaptation results on different splits of the MUSIC-21 dataset (i.e., train on randomly selected 16 music classes from MUSIC-21 and evaluate on the other five classes), we evaluate on the whole MUSIC-21 without using it during training. Our setting is more challenging than SeCo [58] as our method has never seen any multi-source (i.e., duet) videos during training on the MUSIC dataset. Table 4 and 5 summarize the adaptation results on the MUSIC-21 multi-source (solo+duets) and single-source (solo), respectively, under different variants of our approach. Interestingly, cross-modal consistency learning promotes the separation results more aggressively when adapting to unknown music classes than known classes (see Table 7). This result further strengthens the importance of designing meta-consistency learning framework. The performance is further improved (0.45dB SDR) by applying meta-testing. In contrast, applying naive TTA only marginally improves the performance (0.07dB SDR).

4.5. Adaptation Results on AudioSet

We also evaluate our framework on the AudioSet dataset [18] to further demonstrate models ability to separate unknown sound. Table 6 shows the comparison results with other approaches. Interestingly, our method outperforms two out of three methods without any training on this challenging dataset and without any labeled examples to guide test-time adaptation.

Method	Train on AudioSet	SDR ↑	SIR↑
Sound of Pixels [56]	✓	1.66	3.58
AV-MIML [15]	✓	1.83	-
Co-Separation [16]	✓	4.26	7.07
Ours	✗	2.71	7.27

Table 6. Performance comparison on AudioSet test set split provided by [16]. Note that previous methods are both trained and tested on AudioSet. However, ours is only tested on AudioSet.

Method	SDR ↑	SIR↑	SAR↑
AVSS	10.35	16.71	12.43
AVSS + CMC	10.41	17.06	12.13
AVSS + CMC + Meta	10.53	17.51	12.14
AVSS + CMC + Naive TTA	10.42	17.06	12.13
AVSS + CMC + Meta TTA	**10.80**	**17.87**	**12.30**

Table 7. The ablation results comparing different variants of our proposed pipeline on MUSIC val set. Our final method which incorporates both cross-modal consistency and meta-consistency training, outperforms all other baselines by a substantial margin.

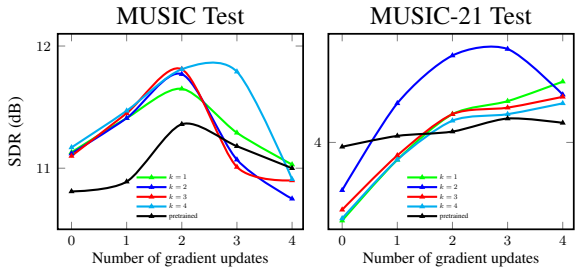

Figure 3. Illustration of SDR after each gradient update for the fully meta-trained models trained with $k = \{1, 2, 3, 4\}$ and the pre-trained consistency model with no meta-training on MUSIC (left) and MUSIC-21 (right) test set. Overall, the meta-consistency learned models improve performance via test time adaptation for all the values of k used in meta-consistency training.

4.6. Ablation Study

4.6.1 Audio-Visual Consistency and Meta Learning

We conduct experiments on MUSIC val split [50] to examine the significance of different components of our method and summarize the results in Table 7. Our baseline AVSS achieves SDR of 10.35dB and 16.71dB of SIR without any audio-level class labels. When we include the cross-modal consistency loss with AVSS, AVSS+CMC marginally outperforms AVSS (10.35dB SDR vs 10.41dB SDR and 16.71dB SIR vs 17.06dB SIR). The meta-consistency training (AVSS+CMC+Meta) further promotes the separation performance marginally (10.53dB SDR and 17.71dB

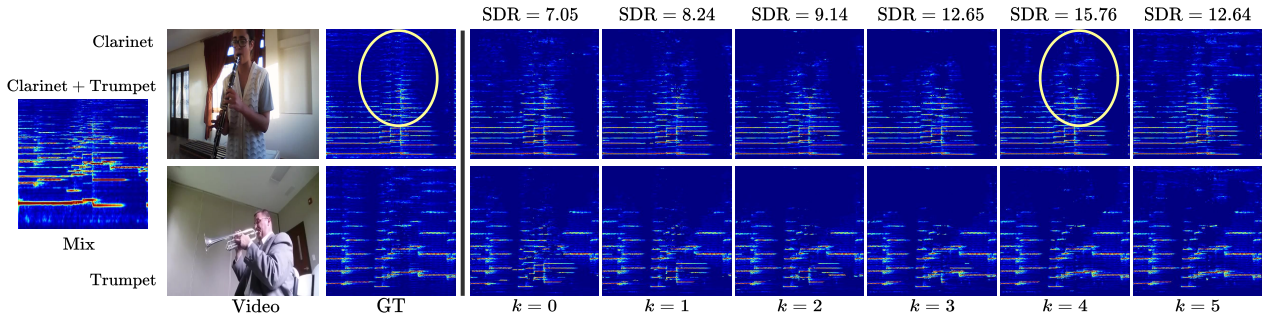

Figure 4. Illustration of unfolded adaptation process for various number of gradient updates during test time adaptation. Interestingly, more cleaner and visually close to ground-truth separated spectrograms are generated with $k = 4$.

SIR). Finally, meta-consistency based test time adaptation (AVSS+CMC+Meta TTA) reasonably improves the overall separation performance (0.27dB SDR improvement). Interestingly, naive test time adaptation (AVSS+CMC+Naive TTA) can not improve separation performance (0.01dB SDR improvement) which reveals the importance of our meta-consistency based training for audio separation task. It is clear that the cross-modal consistency loss and meta-consistency training based test time adaptation promote the separation performance and we achieve the best results by employing both cross-modal consistency loss and meta-consistency training regarding all evaluation metrics.

4.6.2 Effects on the Number of Inner Loop Updates

To analyze the effect of varying number of inner loop updates during meta consistency training on MUSIC val split and test time adaptation on MUSIC test split, we conduct a series of experiments. Figure 3 demonstrates how the overall separation performance changes while varying the inner loop gradient updates from 0 to 5 during meta-testing. We use five different meta-trained models with $k = \{1, 2, 3, 4\}$ and the pretrained consistency model in our experiment settings. In general, test time adaptation with smaller number of inner loop updates (i.e., $k = \{2, 3, 4\}$) shows the most SDR gain, while increasing the number of updates, k does not have any impact on improving separation performance. Note that the results for $k > 4$ is not shown as we do not observe any improvement for those values of k. Interestingly, the separation performance is even diminished with more gradient updates, which is somewhat counter-intuitive of the hypothesis (i.e., larger k allows the model to better adapt to the test sample) compared to the tendency reported in existing works [7, 12, 37]. However, recent work on video frame interpolation [8] achieved best performance with just one gradient update. The possible reasons for this phenomena can be two folds. Firstly, it might cause sever overfitting to the data used for the inner loop updates. In fact, the inner loop may concentrate too much on learning cross-modal consistency and forget the generic prior knowledge learned by baseline pretrained model to separate audios from the mixture. Secondly, the complexity of training grows with the increase of gradient updates, which makes the task harder for the model to find the optimal local minima [12, 33]. Interestingly, naively adapting the pretrained model can not facilitate test time adaptation to improve the separation performance as the meta-consistency learned models. While the existing work [7] pointed out that number of gradient updates should match during meta-auxiliary training and test time adaptation, we found that the separation performance is minimally changed when using different gradient updates during training and test time adaptation. Figure 4 shows the unfolded adaptation process.

5. Discussion and Conclusion

In this paper, we introduced a novel method for visually guided audio source separation. To promote the adaptation ability on known and unknown samples, we designed a cross-modal consistency learning module that learns the synchronization between visual embedding and its corresponding audio embedding to improve separation performance. In addition, we integrated meta consistency learning to constrain the model so that the gradient updates via the consistency loss bring in the performance improvement and promote the ability for faster and effective test time adaptation by quickly adapting its parameters accordingly. Our experimental results demonstrated consistent performance improvement on multiple benchmark datasets. We further conducted extensive ablation studies to emphasize the importance of the key components of our overall method. The main novelty we introduce comes from the overall system design, as opposed to methodological details of individual components. While we re-use certain components from the prior art (*e.g.* multi-modal transformer [28,43], audio-visual consistency [58]), we focus on their novel and optimal combination to effectively support meta-consistency driven test-time adaptation and that way enable quick adaptation of the pretrained model. We envision extensions to a more universal sound separation to include off-screen sounds is an interesting future direction.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019. 2

[3] Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, and Andreas Dengel. Self-supervised test-time adaptation on video data. In *WACV*, 2022. 1, 4

[4] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *ICCV*, 2021. 1, 2, 6

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022. 1, 4

[6] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *CVPR*, 2022. 1, 2, 4, 5

[7] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *CVPR*, 2021. 1, 2, 4, 5, 8

[8] Myungsub Choi, Janghoon Choi, Sungyong Baik, Tae Hyun Kim, and Kyoung Mu Lee. Scene-adaptive video frame interpolation via meta-learning. In *CVPR*, 2020. 1, 2, 4, 5, 8

[9] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. *arXiv preprint arXiv:2005.07074*, 2020. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[11] 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. 2

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1, 2, 8

[13] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017. 2

[14] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020. 1, 2, 3, 6

[15] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 2, 7

[16] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *CVPR*, 2019. 1, 2, 3, 6, 7

[17] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 1, 2, 3, 4

[18] 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 *ICASSP*, 2017. 7

[19] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE TASSP*, 1984. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 4

[21] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *ISMIR*, 2014. 2

[22] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama. Singing voice analysis and editing based on mutually dependent f0 estimation and source separation. In *ICASSP*, 2015. 6

[23] Yanli Ji, Shuo Ma, Xing Xu, Xuelong Li, and Heng Tao Shen. Self-supervised fine-grained cycle-separation network (fscn) for visual-audio separation. *IEEE TMM*, 2022. 1, 2

[24] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *ACCV*, 2018. 4

[25] Sunwoo Kim and Minje Kim. Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation. In *WASPAA*, 2021. 1

[26] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 1, 4

[27] Shikun Liu, Andrew Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. In *NeurIPS*, 2019. 2

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2, 3, 4, 8

[29] Kaiyue Lu, Nick Barnes, Saeed Anwar, and Liang Zheng. From depth what can you see? depth completion via auxiliary image reconstruction. In *CVPR*, 2020. 2

[30] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. 1989. 1, 2

[31] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020. 3

[32] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*, 2018. 1, 4

[33] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 8

[34] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022. 1

[35] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2

[36] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. Motion informed audio source separation. In *ICASSP*, 2017. 1, 2

[37] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *CVPR*, 2019. 1, 2, 4, 5, 8

[38] Seobin Park, Jinsu Yoo, Donghyeon Cho, Jiwon Kim, and Tae Hyun Kim. Fast adaptation to super-resolution networks via meta-learning. In *ECCV*, 2020. 2

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[40] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014. 6

[41] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *CVPR*, 2022. 2, 3, 4

[42] Tanzila Rahman and Leonid Sigal. Weakly-supervised audio-visual sound source detection and separation. In *ICME*, 2021. 6

[43] Tanzila Rahman, Mengyu Yang, and Leonid Sigal. Tribert: Human-centric audio-visual representation learning. In *NeurIPS*, 2021. 1, 2, 3, 4, 5, 6, 8

[44] 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*, 2015. 3

[45] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised segmentation and source separation on videos. In *CVPR Workshops*, 2019. 2

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3, 4

[47] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *ICDAE*, 2009. 6

[48] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 1, 4

[49] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2019. 3

[50] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[52] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 4

[53] Yong Wu, Shekhor Chanda, Mehrdad Hosseinzadeh, Zhi Liu, and Yang Wang. Few-shot learning of compact models via task-specific meta distillation. In *WACV*, 2022. 4, 5

[54] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *ICCV*, 2019. 1, 2, 6

[55] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 1, 2, 3, 6

[56] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7

[57] Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. In *NeurIPS*, 2022. 4, 5

[58] Xinchi Zhou, Dongzhan Zhou, Wanli Ouyang, Hang Zhou, Ziwei Liu, and Di Hu. Seco: Separating unknown musical visual sounds with consistency guidance. *arXiv preprint arXiv:2203.13535*, 2022. 1, 2, 3, 4, 7, 8

[59] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation using cascaded opponent filter network. In *ACCV*, 2020. 2

[60] Lingyu Zhu and Esa Rahtu. Leveraging category information for single-frame visual sound source separation. In *EUVIP*, 2021. 2

[61] Lingyu Zhu and Esa Rahtu. V-slowfast network for efficient visual sound separation. In *WACV*, 2022. 2

[62] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. In *CVPR*, 2022. 1, 2, 3, 5