This WACV 2023 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SeCo: Separating Unknown Musical Visual Sounds with Consistency Guidance

Xinchi Zhou[1] Dongzhan Zhou[1] Wanli Ouyang[1] Hang Zhou[2] Di Hu[3]
[1]The University of Sydney [2]Baidu Inc.
[3]Gaoling School of Artificial Intelligence, Renmin University of China
{xinchi.zhou1,d.zhou,wanli.ouyang}@sydney.edu.au,
zhouhang09@baidu.com, dihu@ruc.edu.cn

Abstract

Recent years have witnessed the success of deep learning on the visual sound separation task. However, existing works follow similar settings where the training and testing datasets share the same musical instrument categories, which to some extent limits the versatility of this task. In this work, we focus on a more general and challenging scenario, namely the separation of unknown musical instruments, where the categories in training and testing phases have no direct overlap with each other. To tackle this new setting, we propose the "Separation-with-Consistency" (SeCo) framework, which can accomplish the separation on unknown categories by exploiting the consistency constraints. Furthermore, to capture richer characteristics of the novel melodies, we devise an online matching strategy, which can bring stable enhancements with no cost of extra parameters. Experiments demonstrate that our SeCo framework exhibits strong adaptation ability on the novel musical categories and outperforms the baseline methods by a notable margin.

1. Introduction

The objective of visual sound separation is to separate the mixed audio signals into individual components with the guidance of visual cues. Deep neural networks can extract rich semantic information from both visual and auditory modalities, which significantly promote the development of the visual sound separation task. Most deep-learning based approaches, such as [44, 13, 11], adopt the setting where the training and testing sets share the same musical instrument categories. Despite the success, there still exist some limitations on such training mode, which confine the separation targets to the musical instruments that have appeared in the training set. The more general setting of visual sound separation on unknown musical instruments remains an unexplored problem.

In this work, we undertake the task of visually guided music separation on unknown classes, that is, the categories of training and testing sets have ***no direct overlap***. Under the real scenario, it is challenging to directly identify and separate the sound of the unfamiliar musical instruments from the mixed audio signal, even for humans. Thus, it may be difficult to directly apply the existing frameworks to this new setting and reasonable priors must be added to enhance the adaptability of the deep models to the unknown musical sounds.

To handle this challenging new setting, we propose a novel 'Separation with Consistency' (SeCo) framework, which exploits the consistency constraints to realize the visual sound separation on the novel musical instruments. The system receives two types of consistency supervisions, namely the **inter-modal** consistency and the **intra-modal** consistency. Firstly, the audio-visual associations in videos are natural and will not change with different categories. Therefore, it is critical to strengthen the audio-visual (AV) correlations during training, instead of simply capturing the isolated features for the auditory and visual modalities. In this way, the visual cues can provide better separation guidance even for the categories that have never been seen before. Specifically, we require that the separated audio signals should be aligned with the visual components in the original videos, which is denoted as the inter-modal consistency. Secondly, even though directly identifying unfamiliar sounds is not easy for humans, things will become quite different if some auditory examples are provided. In particular, human brains can achieve the goal of separating the novel target sound by perceiving the similarities and differences between the mixed sound and the given template sound. Thus, it is a natural idea to incorporate such a template learning mechanism in deep models for the assistance of visual sound separation on unknown classes. Specifically, since the sounds from the same type of musical instruments normally enjoy similar timbres and tones, the features extracted from them should be close in the embedding space. In this way, the intra-modal consistency expands the supervision scale from the sample level to the wider category level, which can effectively help the deep models adapt

well to the unknown classes.

For an unfamiliar melody, humans may listen to the melody over and over again to better capture the characteristics of the musical tone. Similarly, this behavior can also be applied to our task. We develop an online matching strategy to iteratively refine the predicted mask for each sample independently so that the potential of the devised consistency guidance can be further exploited. The online matching strategy can bring stable improvements without introducing any extra parameters.

Our **contributions** can be summarized as followed. (1) We explore the task of visual music separation under the scene of unknown musical instruments, which expands the scope of visual sound separation and makes the task more versatile. (2) We propose a novel framework, SeCo, to adapt to this challenging situation. The results show that our approach outperforms the baselines by a noticeable margin. We also conduct in-depth ablation studies to analyze the effects of the key parameters. (3) We design an online matching strategy, which brings consistent improvements with no extra parameter costs.

2. Related Work

Audio-Visual Learning. With the development of deep learning, audio-visual learning has also received widespread attention in recent years and breakthroughs have been made in various sub-fields [47, 41]. Audio-visual representation learning aims at finding the correlations between the audio and visual modality in a self-supervised manner and thus provides good audio/visual representations [1, 2, 30, 23, 29, 17]. In addition, many works utilize audio information to improve the video analysis tasks [20, 15]. The objective of the audio-visual localization task is to localize the sound source in the visual context [34, 44, 2, 18]. Another important branch of the audio-visual learning field lies in the cross-modality generation, which consists of visual-to-audio [46, 27] and audio-to-visual [5, 4, 16, 45] tasks. Most previous works in audio-visual learning require that the training and validation data come from the same domain or similar scenario, while our work investigates a more challenging setting where the training and testing set has no direct category overlaps.

Visual Sound Separation. By leveraging visual modality to the sound separation task, models can utilize the richer context information, which outperforms the single modality approaches. Visual sound separation is explored on various identities, such as speakers [10, 6, 26], objects [12] and musical sounds [44, 42, 13, 42, 43, 11]. Our work concentrates on the branch of visual music separation.

Zhao et al. [44] propose the PixelPlayer framework with the 'mix-and-separation' paradigm, which learns to separate mixed audios into components and locate the sound production regions on images in a self-supervised manner. Considering the limitations of static images, many works attempt to adopt visual cues from other modalities and further benefit the separation task, such as motion [43], skeleton [11] and scene graph [3]. Gao et al. [14] incorporate audio-visual consistency in the speech separation framework but the training and validation sets act on the same category, i.e., speakers. Unlike any of the above, we focus on the visual sound separation task of unknown musical instruments and also propose an effective framework to handle this new scene.

Transfer Learning on Novel Category. Humans naturally have a strong ability to establish the perception of new objects, even with very limited samples. However, in most cases, machines can obtain such perception ability only if it has been fed enough examples. Thus, few-shot [7] and zero-shot [24] learning are proposed to investigate the transferability of machines when very few or even no samples are provided on new objects. Such learning paradigms can effectively reduce the burdens of data acquisition and storage. Few-shot learning approaches utilize the information of the limited samples from the new category, while the models have no exposure to any instances of the target class under the zero-shot setting. The mainstream solutions for the few-shot setting include metric learning [21, 35, 37] and meta-learning [9, 28]. Zero-shot learning approaches usually transfer knowledge from familiar classes, such as semantic embedding [25, 22], or exploit external information such as knowledge graphs [40]. In addition to the classification scene, many works extend the few/zero-shot setting to other sub-fields such as object detection [19] and semantic segmentation [39]. Despite the success in vision fields, it is still challenging to deploy novel category transfer learning on multi-modality models. By leveraging the inter-modal and intra-modal consistency guidance, our SeCo framework exhibits impressive transferring performance on unknown musical separation and serves as a strong baseline for this novel and challenging task.

3. Methodology

We propose the 'Separation-with-Consistency' paradigm (SeCo) to achieve the transfer learning of visual sound separation on novel musical instruments. Specifically, the consistency guidance is composed of the inter-modal and intra-modal parts, which require the separated sounds to align with the corresponding visual cues and sounds of the same category. The pipeline of our SeCo framework is illustrated in Fig. 1.

3.1. Framework Overview

The goal of the visual sound separation task is to separate the sound components from the mixed signal by leveraging the visual information. Following previous works [44, 13, 43], we also adopt the 'mix-and-separation' paradigm to carry out the training in a self-supervised manner.

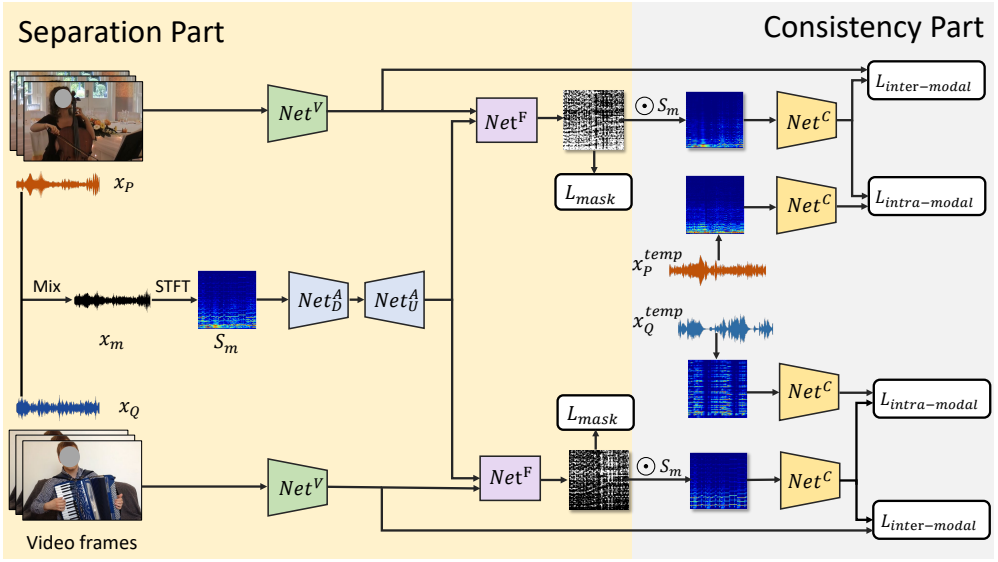

Figure 1. The whole pipeline of our "Separation-with-Consistency" framework is composed of the separation part and the consistency part. In the separation stage, the visual features and audio features are extracted by the vision network Net^V and audio network Net^A, respectively, and get fused in the fusion network Net^F to predict the separation masks. In the consistency stage, the separated spectrograms and the template spectrograms pass through the consistency network Net^C to generate the high-level features for the computation of the consistency constraints. The system is trained by minimizing the combination of the separation loss (L_{mask}) and the consistency loss ($L_{intra-modal}$ & $L_{inter-modal}$).

Suppose we have two video clips $\{P, Q\}$ with corresponding audio signals $\{x_P, x_Q\}$, the audio components are mixed to generate a synthetic mixture signal $x_m = x_P + x_Q$. For easy training, the mixed raw signal x_m is first converted to the spectrogram $\mathbf{S_m}$ via Short Time Fourier Transform (STFT). The vision analysis network extracts the visual feature $\mathbf{f_i^v}$ ($i \in \{P, Q\}$) from the input frames for each video clip, while the audio feature $\mathbf{f_a}$ is generated by feeding the mixed spectrogram $\mathbf{S_m}$ into the audio network. Afterward, the audio feature is fused with the visual features $\{\mathbf{f_P^v}, \mathbf{f_Q^v}\}$, respectively, to produce the separation masks $\{\mathbf{M_P}, \mathbf{M_Q}\}$. Finally, we multiply the mixed spectrogram with the predicted masks to obtain the clean spectrograms and produce the clean signals via Inverse STFT.

Different from the original setting, we aim to explore a more challenging scenario to separate the unknown musical instruments. To achieve the adaptation ability on the novel categories, we introduce an additional consistency analysis network, which requires the predicted separation results to maintain both the **inter-modal** and **intra-modal** consistency. The inter-modal consistency is implemented with the synchronization of video and the corresponding separated audio [23, 29], where the network can capture the audio-visual correlations when encountering new categories and acquire stronger transferring ability. Besides, based on the observation that instruments of the same type normally have similar timbres and tones, we add the intra-modal consistency supervision to the system, which will shorten the distance of the audio features from the same category and enlarge that from different categories in the embedding space.

Inspired by the fact that humans may spend more time observing and exploring repeatedly when encountering unfamiliar objects, we introduce the online matching mechanism during the inference stage so that the model can better fit the new instrument category. Specifically, the framework will make explicit adjustments for each sample pair by recurrently updating the model parameters from the supervision of the consistency loss. Please note that *no* Ground-Truth masks are required since we only adopt the consistency loss as the error signal. The initial separation signals may be coarse due to the considerable gap between the training domain and the testing domain. But as the model becomes more familiar with the test sample, it can grasp more precise information and hence generate more delicate audio components.

3.2. Separation Network

The separation network is composed of three components, that is, the vision analysis network, the audio network, and a fusion network. The mixed spectrogram passes through the audio network to generate the audio feature. The visual feature is extracted by the vision analysis network for each video clip and then fused with the mixed audio feature in the fusion network to produce the separation mask. The process is illustrated in Fig. 2.

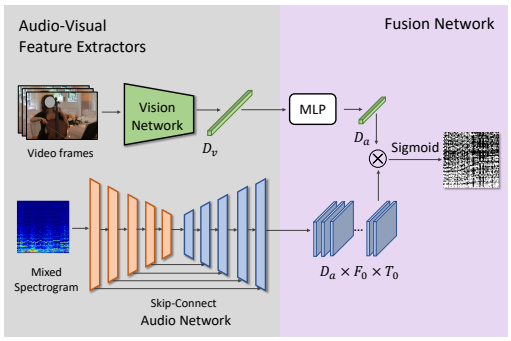

Figure 2. Structure of the separation network.

3.2.1 Vision analysis network

Videos contain rich visual cues, such as appearances, texture, motion, and so on. In our framework, we focus on motion information as the visual message for the following considerations. First, compared with the spatial semantics, motion relies less on category information, which makes it more effective guidance when dealing with novel classes. Second, humans can naturally associate the instrument playing actions with the sounds, regardless of the specific instrument category. Thus, exploiting such correlation will enhance the transferring ability of the system.

We adopt the fast pathway in the SlowFast network [8] as our vision analysis network to extract visual features. As the motion information is selected as the primary visual guidance, we remove the slow pathway and the lateral connection structure from the original SlowFast architecture and only keep the fast branch. Our vision network also preserves the high temporal resolution and low channel capacity properties, which can capture the detailed motions without introducing heavy parameter burdens. The specific implementation of the network will be provided in the supplementary. The vision network does not need optical flows and can directly learn the motion representations from the raw frames in an end-to-end manner.

3.2.2 Audio network

Following previous works [44, 13], we adopt a U-Net [33] style encoder-decoder with skip-connections to extract the audio features. The U-Net consists of 5 downsampling convolution layers and 5 de-convolution layers for upsampling. The audio network takes the mixed spectrogram $\mathbf{S_m}$ as input and yields audio feature of shape $D_a \times T_0 \times F_0$, where T_0 and F_0 refer to the temporal and frequency dimensions, respectively, and have the same values as $\mathbf{S_m}$. If not specified, we set $D_a = 64$ in the experiments.

3.2.3 Fusion network

After the visual and audio features are extracted, we can fuse the visual guidance into the audio feature to compute the separation mask. Before the fusion, we adjust the channel dimension of the visual feature to D_a via a linear projection and then apply a sigmoid activation on the projected feature. The activated visual feature is multiplied with the audio feature along the channel dimension to compute the fusion mask of shape $1 \times T_0 \times F_0$. Finally, we activate the fusion mask via the sigmoid function to acquire the predicted separation mask. The separation loss L_{mask} is the per-pixel binary cross-entropy loss between the predicted mask and the Ground-Truth mask. The Ground-Truth mask of each component is produced by checking whether the target spectrogram is dominant in the mixed spectrogram at every T-F unit:

$$M_i^{GT}(x,y) = [S_i(x,y) \geq S_m(x,y)], i \in \{P, Q\}, \quad (1)$$

where (x, y) refers to the coordinates along the T-F dimensions.

3.3. Consistency Network

To raise the adaptation ability on novel classes, we propose two types of consistency contraints, i.e., inter-modal consistency and intra-modal consistency, which are both exerted on the predicted separation results. Thus, the mixed spectrogram $\mathbf{S_m}$ is multiplied by the predicted masks $\{\mathbf{M_P}, \mathbf{M_Q}\}$ to develop the separated spectrograms $\{\mathbf{S_P^{pred}}, \mathbf{S_Q^{pred}}\}$. Furthermore, since the comparison of the raw spectrograms may not be very informative, we use high-level features to replace the low-level spectrograms for consistency computation, which are extracted from the consistency network. The network is stacked by 10 residual blocks, followed by a global max-pooling layer. The consistency embeddings of $\{\mathbf{S_P^{pred}}, \mathbf{S_Q^{pred}}\}$ are denoted as $\{\mathbf{f_P^{pred}}, \mathbf{f_Q^{pred}}\}$, respectively.

3.3.1 Inter-modal consistency

Since the audio-visual associations in videos are natural and will not be disturbed by the category information, we add the inter-modal consistency to the system to strengthen the synchronization between the audio and visual elements so that it will present stronger adaptation ability with novel categories. Similar to [23], the training objective is minimizing the distance on the positive pairs while enlarging the distance on the negative pairs. The positive pairs are synchronized audio-visual samples, i.e., the separated audio embeddings and their corresponding visual features $\{\mathbf{f_i^{pred}}, \mathbf{f_i^v}\}$, $i \in \{P, Q\}$. The negative pairs are obtained by cross-pairing the uncorrelated audio and visual features, that is, $\{\mathbf{f_i^{pred}}, \mathbf{f_j^v}\}, i \neq j, i, j \in \{P, Q\}$.

At the beginning of training, the separation results may be poor since the network has not fully converged yet, and the suboptimal separation predictions may confuse the identification of positive pairs. Based on this consideration, we

introduce the Ground-Truth audio features to assist the synchronization learning, which are extracted from $\{\mathbf{S_P}, \mathbf{S_Q}\}$ and denoted as $\{\mathbf{f_P^{GT}}, \mathbf{f_Q^{GT}}\}$. The loss weights between the predicted part and the Ground-Truth part vary according to the training time. The inter-modal consistency loss is defined as follows:

$$\begin{aligned} L_{inter-modal} = \gamma(t)(&D(\mathbf{f_P^{GT}}, \mathbf{f_P^v}) + D(\mathbf{f_Q^{GT}}, \mathbf{f_Q^v})) \\ &+ D(\mathbf{f_P^{pred}}, \mathbf{f_P^v}) + D(\mathbf{f_Q^{pred}}, \mathbf{f_Q^v}) \\ &- D(\mathbf{f_P^{pred}}, \mathbf{f_Q^v}) - D(\mathbf{f_Q^{pred}}, \mathbf{f_P^v}), \end{aligned} \tag{2}$$

where D refers to the L_2 distance between two features and $\gamma(t)$ is the weight for the Ground-Truth assisted part that decays over training time. All features are normalized before computation.

3.3.2 Intra-modal consistency

The design of the intra-modal consistency is based on two assumptions: (1) Instruments of the same category should have similar tones and timbres so their audio signals are supposed to be closer when projected to the feature space. (2) To achieve a higher quality separation result, the audio features of the two mixed videos should be pulled away. Please note that assumption (2) does not conflict with (1) because we require that the two mixed audios come from different instrument classes.

For the in-class similarity learning in assumption (1), we utilize audio signals $\{x_P^{temp}, x_Q^{temp}\}$ from the additionally sampled template video clips. Please note that the template clip comes from a different video of the same category as the separation target. The template audio signals are also converted to spectrograms via STFT and then pass through the consistency network to produce the high-level embeddings $\{\mathbf{f_P^{temp}}, \mathbf{f_Q^{temp}}\}$. The intra-modal consistency loss is shown as followed:

$$\begin{aligned} L_{intra-modal} = &D(\mathbf{f_P^{pred}}, \mathbf{f_P^{temp}}) + D(\mathbf{f_Q^{pred}}, \mathbf{f_Q^{temp}}) \\ &- D(\mathbf{f_P^{pred}}, \mathbf{f_Q^{pred}}), \end{aligned} \tag{3}$$

where D represents the L_2 distance between the features and all features are normalized before computation. The consistency loss L_{cs} is the sum of the inter-modal and intra-modal components:

$$L_{cs} = L_{inter-modal} + L_{intra-modal} \tag{4}$$

Therefore, the overall loss function of our framework is:

$$L = L_{mask} + \lambda L_{cs}, \tag{5}$$

where λ is the weight of consistency loss.

3.4. Online Matching Strategy

We introduce an online matching strategy to promote model compatibility with the samples from the novel domain in the inference phase. The parameters of networks will be fine-tuned explicitly for each sample pair by the backpropagation of error signals from the consistency loss. In this way, the online matching process can be regarded as 'training during inference' but we only adopt the consistency loss as the supervision signal, and the optimization is based on one single pair. We emphasize that *no* Ground-Truth separation masks are involved in this process so that the process can be regarded as a self-correction mechanism (the Ground-Truth assisted part in $L_{inter-modal}$ is excluded).

For each sample pair, we optimize the model parameters via the consistency loss for several iterations and generate the refined separation masks for the pair based on the updated parameters. Before moving to the next pair, the parameters are switched to the original state so that the samples will not mutually affect each other. In practice, we fix all BatchNorm layers to avoid the fluctuations caused by the single sample input. Please refer to the supplementary for more details about the process. Our online matching strategy will not introduce any extra parameters but can bring consistent improvements.

4. Experiments

4.1. Implementation Details

Our pipeline is implemented with the PyTorch framework [31]. We use an Adam optimizer with betas (0.9, 0.999) and batch size 40. The weight of consistency loss λ in Eq. 5 is set to 0.01. The decay parameter $\gamma(t)$ in Eq. 2 follows the function: $\gamma(t)$ = max(0.1, $0.9^{iter/100}$), where *iter* refers to the training iterations. The framework is trained for 17000 iterations. The learning rate of the vision and consistency network are 1e-4 while the that of the audio and fusion network are 1e-3. During the online matching process, the learning rate is 1e-4 for the entire system and each sample pair is refined for 5 iterations. Since the consistency loss is the only supervision signal, we set λ to 1.0 in this process. The data processing details will be provided in the supplementary.

4.2. Dataset and Evaluation Metrics

We quantitatively evaluate our framework on the MUSIC-21 dataset [43] which contains 21 classes of instruments. The dataset is composed of untrimmed videos crawled from the YouTube website so that the contents are relatively diverse and complex. We randomly select 16 instruments as the training split and use the other 5 classes as the testing split, denoted as split-1. More details are provided in the supplementary.

Method	SDR	SIR	SAR
NMF-MFCC [36]	0.90	5.37	6.94
Sound-of-Pixels [44]	-2.56	2.42	4.97
Co-Separation [13]	-2.89	1.97	5.23
MPNet [42]	-2.32	2.07	5.54
Sound-of-Motions [43]	0.81	3.44	7.06
SeCo (motion only)	1.16	4.39	9.64
SeCo	4.01	7.13	11.62

Table 1. Sound separation results on the MUSIC-21 testing dataset, higher is better for all metrics. The SeCo (motion only) does not adopt the consistency loss and utilizes the motion information as visual guidance. SeCo incorporates both the consistency loss and the online matching strategy, which outperforms all baselines by a large margin. The results are reproduced with official codes as existing pre-trained models are trained on all categories in MUSIC-21.

We use the open-source mir_eval library [32] to conduct quantitative evaluations on the separated audios, where three metrics are selected: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). The units are dB. The SDR score is normally regarded as the most convincing metric.

4.3. Quantitative Results

The results of the baseline and our SeCo are shown in Table 1. The traditional method NMF-MFCC [36] does not exhibit obvious degeneration on the testing splits, which is reasonable since it is non-learned. The traditional algorithm can only return unpaired separation signals so that we conduct the exclusive matchings and take the overall best results. Even with the best matching, it still presents trivial performances, which means that the traditional method lacks the potential for further improvements. The deep-learning based Sound-of-Pixels [44], Co-separation [13], and MPNet [42] methods only adopt the spatial semantics as visual cues and do not explicitly capture the temporal correlations. The results indicate that they fail to successfully separate the sounds of the novel classes. The spatial semantics such as appearances and textures are closely related to the category so the learned visual representations cannot provide sufficient separation guidance when encountering the novel classes.

On the other hand, if the basic components of the visual guidance are transferred from spatial to temporal, i.e., the motion information, the over-fitting symptom can be alleviated, where we can see that motion only SeCo provides a relatively good baseline. Compared with the spatial semantics, the temporal information is less category-specific, which can serve as more effective separation guidance with the novel instrument types. Explanations for the effectiveness may come from the following two aspects. Firstly, the temporal information of motions can better interact with the audio signals since there exists a natural correspondence between the player's movements and sound components. Exploiting such correlations will reduce the dependence on categories and enhance the adaptation ability to new instruments. Secondly, [43, 48] show that motion cues can be used to guide the separation on duets of the same instruments while [44] exhibits inferior performances, which also indicates that motions are less dependent on categories.

$L_{inter-modal}$	$L_{intra-modal}$	SDR	SIR	SAR
✗	✗	1.16	4.39	9.64
✓	✗	2.05	4.93	10.48
✗	✓	2.16	**5.40**	10.01
✓	✓	**2.37**	5.03	**11.29**

Table 2. Ablation study of the importance of consistency loss components on the MUSIC-21 testing dataset.

We also compare with [43], which adopts optical flow as the visual guidance. The results indicate that explicitly using motion information may also benefit the separation process of novel categories. However, it is still inferior to directly learning the motion representations from the raw frames, probably due to the noise in optical flow estimations. Considering the additional computational cost to extract optical flow, learning motion representations in an end-to-end manner may be a more reasonable choice.

Despite the progress of motion representations, simply replacing the visual modality still fails to bring satisfying performance improvements. In general, the baseline results demonstrate that it is a challenge for the normal frameworks to handle the sound separation task on musical instruments that have never seen before. Our SeCo framework outperforms all baseline methods by a large margin under this challenging scenario, which demonstrates the effectiveness of our method. Although changing the visual modality can bring a relatively good starting point, we argue that the main improvements come from the consistency loss. By accomplishing the music separation on novel categories, our method outcomes the limitation of prior works and proves the feasibility of deploying a more general setting. The results may expand the scope of visual music separation and make the task more versatile.

4.4. Ablation Study

4.4.1 Inter-modal v.s. intra-modal consistency

We conduct experiments to investigate the importance of the different loss components and report the results in Table 2. We can see that both the inter-modal consistency and the intra-modal consistency will promote the separation performance on novel musical instrument types, since adopting either loss will win the baseline method (*w.o.* the consistency loss). As depicted in the Table, we achieve the best results by employing both inter-modal and intra-modal consistency losses regarding all evaluation criteria.

	Baseline			SeCo (*w.* O.M.)		
	SDR	SIR	SAR	SDR	SIR	SAR
Image	-2.56	2.42	4.97	2.95	6.34	9.45
Skeleton	-1.35	2.83	6.05	3.43	**7.82**	9.97
Motion	**1.16**	**4.39**	**11.10**	**3.91**	6.50	**11.34**

Table 3. Sound separation results when utilizing visual cues of different modalities. We report results from both the baseline and the SeCo method. The SeCo method includes the normal training and the subsequent online matching process.

4.4.2 Comparison of different visual cues

To investigate the effects of visual cues, we conduct experiments on three visual modalities, i.e., image, skeleton, and motion. Implementation details of the image and skeleton are provided in the supplementary.

The performance comparisons of different visual modalities are summarized in Table 3, where both the baseline and our SeCo approaches are presented. From the baseline results, we can see that when using the static images as the visual guidance, the model fails to successfully separate sounds from the unknown musical instruments. We suspect that the failure may come from the dependence of spatial information on categories, which causes over-fitting to the training scenarios, as analyzed in Sec. 4.3. The visual features of the skeleton modality incorporate both the spatial and temporal relations and we can see that the over-fitting problem has been a little bit alleviated. However, simply replacing the images with skeletons is not enough to generate the optimal results. The possible reason is that the skeleton data only retain the joint coordinates of the players while discarding much detailed information in the original video clips. Such simple and intuitive visual cues may hinder the ability to conduct visual sound separation on new categories, given no additional prior knowledge. In contrast, for the motion modality, the 3d-CNN based vision analysis network directly learns the temporal representations from the original video clips, which can capture richer semantics. This property makes the motion modality better visual cues in our setting and provides an advanced starting point for further improvements.

In addition to the analysis of the baseline results, we also evaluate the performances of our SeCo framework when utilizing different vision modalities. Please note that the inter-modality loss is not applicable to the image-based visual cues. Therefore, for a fair comparison, only the intra-modality loss is applied as the consistency constraints for all modalities. The SeCo pipeline includes both the normal training and the online matching process and we can find that SeCo considerably exceeds the baseline on all three modalities. The results verify the robustness and flexibility of our approach.

4.4.3 Division of musical instrument categories

To ensure that our SeCo framework does not rely on certain instrument types, we make verifications on different train/test splits. These extra train/test splits also follow the 16/5 category division but the internal instrument types vary from each other. We conduct experiments on 2 additional splits and list the results in Table 4, which demonstrate that our SeCo framework is effective on various splits rather than constrained to certain specific instrument types. The category divisions of the splits are provided in the supplementary. Moreover, the results also confirm the robustness of our online matching strategy, which can also handle different instrument types.

	w/o O.M.			*w.* O.M.		
	SDR	SIR	SAR	SDR	SIR	SAR
Split-1	2.37	5.03	11.29	4.01	7.13	11.62
Split-2	3.81	6.28	12.67	5.72	8.87	13.38
Split-3	2.72	5.23	12.04	3.89	6.93	12.50

Table 4. Separation results on different train/test splits. We show results *w/o* and *w.* the online matching strategy (denoted as O.M. in table), respectively.

4.4.4 Online matching iterations

The key hyper-parameter of the online matching strategy is the number of optimization steps at each sample pair, denoted as *iterations*. We investigate the influence of changing the optimization iterations and visualize the trends in Fig. 4. We use the SDR scores to represent the performances since it is the most important metric. Naturally, increasing the iterations will help the network get more familiar with the current sample and thus produce better separation results but we can also observe the marginal effect with longer iterations. The trend also indicates that the online matching strategy can bring stable improvements instead of random fluctuations. Please note that the online matching strategy will only update the existing parameters so that the performance gains are obtained at no cost of extra parameters.

4.5. Qualitative Results

We visualize four cases of the separated spectrograms on the MUSIC-21 testing dataset in Fig. 3. In (a) and (b), we compare the baseline method and the SeCo method. Both methods adopt the motion information as visual cues but the consistency loss is not included in the baseline method. We can observe that the baseline results lose many details and contain components from its mixture audio counterpart, while the SeCo results are closer to the Ground-Truth spectrograms. The comparisons vividly show the effectiveness of the consistency loss.

Although our SeCo method is superior to the baseline method, it may still encounter the detail missing and noisy

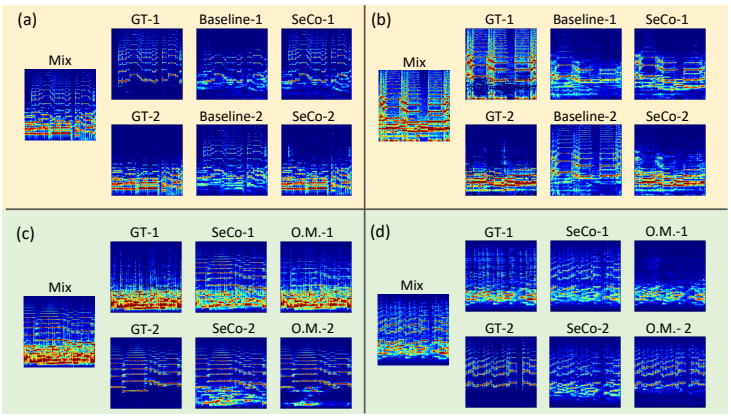

Figure 3. Visualization of the separated spectrograms on the MUSIC-21 testing dataset. The index '1 & 2' refers to the two audio components to be separated and GT stands for the 'Ground-Truth' spectrograms.

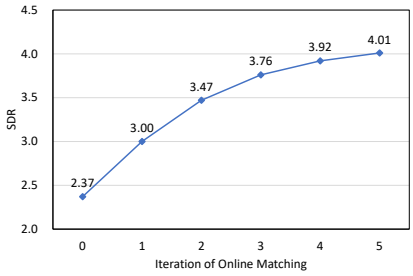

Figure 4. Trend of the SDR scores with different online matching iterations. Iteration 0 refers to the result of not adopting the online matching strategy.

problems due to the challenge of the unknown musical sound separation task. However, these problems can be alleviated by the subsequent online matching process. As shown in (c) and (d), the online matching process can correct the undesirable effects from the other audio component and grasp more details. In this way, we can obtain separation results of higher quality.

5. Discussion and Future Work

Experimental results demonstrate that existing visual sound separation frameworks do not inherently possess the ability to generalize well on novel instrument categories. As a preliminary work, we leverage some priors (e.g., templates from the same category) to enhance the transferability of unseen instruments during the training and testing stages, which raises the separation performance and verifies the feasibility of this setting. Compared with the delicate point-wise separation masks, categories can be regarded as global messages, and we find that such coarse priors can assist the process. However, we hope those priors can be removed in future explorations to improve the versatility of this setting further.

To explore the effects of the visual analysis network, we change the backbone from FastNet to R3D [38] and observe the SDR score decreases to 2.51 dB. The possible reason is that R3D pays less attention to the motion messages, affecting its adaptation to new instrument categories. The results also indicate that visual encoders play an important role in the separation process, especially in this setting. Although FastNet is an economical solution, we wish to see more explorations on the visual encoders to provide more effective separation guidance.

6. Conclusions

In this work, we explore a novel and challenging scenario of visual sound separation, i.e., music separation on unknown musical instruments. To promote the adaptation ability for the deep model on unfamiliar melodies, we design the Separation-with-Consistency (SeCo) framework that utilizes both the inter-modal and intra-modal consistency constraints. Moreover, to fully exploit the consistency potentials, we devise the online matching strategy, which further boosts the system performance with no extra parameter costs. We conduct extensive ablation studies to analyze the key factors in the system, which also exhibit that our SeCo framework is effective and robust on various visual modalities and musical instrument types. Our work proves the feasibility of separation on novel musical instruments and hence expands the scope of the visual sound separation task. We wish our work could inspire the community to further explore the transferability of deep models in the audio-visual learning field.

Acknowledgement

Wanli Ouyang was supported by the Australian Research Council Grant DP200103223, Australian Medical Research Future Fund MRFAI000085, CRC-P Smart Material Recovery Facility (SMRF) – Curby Soft Plastics, and CRC-P ARIA - Bionic Visual-Spatial Prosthesis for the Blind. Di Hu was supported by the National Natural Science Foundation of China (NO.62106272) and the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001).

References

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

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

[3] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021.

[4] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[5] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017.

[6] 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.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[8] 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.

[9] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *arXiv preprint arXiv:1710.11622*, 2017.

[10] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through noise: Visually driven speaker separation and enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3051–3055. IEEE, 2018.

[11] 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.

[12] 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.

[13] 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.

[14] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021.

[15] 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.

[16] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

[17] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019.

[18] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[20] 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.

[21] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[22] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.

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

[24] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[25] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.

[26] Rui Lu, Zhiyao Duan, and Changshui Zhang. Listen and look: Audio–visual matching assisted speech source separation. *IEEE Signal Processing Letters*, 25(9):1315–1319, 2018.

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

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

[29] 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)*, pages 631–648, 2018.

[30] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *International Journal of Computer Vision*, 126(10):1120–1137, 2018.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[32] 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 *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[33] 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.

[34] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[36] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of the 12th International Conference on Digital Audio Effects*, 2009.

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[39] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019.

[40] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018.

[41] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.

[42] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–891, 2019.

[43] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019.

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

[45] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019.

[46] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018.

[47] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, pages 1–26, 2021.

[48] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1289–1299, 2022.