SepFusion: Finding Optimal Fusion Structures for Visual Sound Separation

Dongzhan Zhou[1]*, Xinchi Zhou[1]*, Di Hu[2]†, Hang Zhou[3], Lei Bai[1], Ziwei Liu[4], Wanli Ouyang[1]

[1]The University of Sydney,
[2]Gaoling School of Artificial Intelligence, Renmin University of China,
[3]Baidu Inc.,
[4]S-lab, Nanyang Technological University
{d.zhou, xinchi.zhou1, lei.bai, wanli.ouyang}@sydney.edu.au,
dihu@ruc.edu.cn, zhouhang09@baidu.com, ziwei.liu@ntu.edu.sg

Abstract

Multiple modalities can provide rich semantic information; and exploiting such information will normally lead to better performance compared with the single-modality counterpart. However, it is not easy to devise an effective cross-modal fusion structure due to the variations of feature dimensions and semantics, especially when the inputs even come from different sensors, as in the field of audio-visual learning. In this work, we propose SepFusion, a novel framework that can smoothly produce optimal fusion structures for visual-sound separation. The framework is composed of two components, namely the model generator and the evaluator. To construct the generator, we devise a lightweight architecture space that can adapt to different input modalities. In this way, we can easily obtain audio-visual fusion structures according to our demands. For the evaluator, we adopt the idea of neural architecture search to select superior networks effectively. This automatic process can significantly save human efforts while achieving competitive performances. Moreover, since our SepFusion provides a series of strong models, we can utilize the model family for broader applications, such as further promoting performance via model assembly, or providing suitable architectures for the separation of certain instrument classes. These potential applications further enhance the competitiveness of our approach.

Introduction

Recent years have witnessed the great development of audio-visual learning, where deep neural networks perform an important role as feature encoders and decoders. Such intelligent systems could perceive the world synthetically by processing various input information from multiple modalities simultaneously, which have benefited broad applications in many fields such as sound recognition, music source separation, stereo sound generation, and so on.

Visual cues play an important role in the audio analysis system, especially for the visual sound separation task. For example, it can provide knowledge about object appearances and textures to guide the sound separation process (Zhao et al. 2018; Gao and Grauman 2019b; Zhao et al. 2019; Ephrat et al. 2018). One of the most important issues in visual sound separation is how to effectively fuse vision information into the audio backbone. However, the cross-modality fusion is never easy due to the complex feature structure and diverse semantic context, *e.g.* it is hard to combine a visual feature whose pixels have spatial relations with an audio feature containing frequency-related information. While previous methods have proposed different fusion techniques with diverse visual modalities (Gao and Grauman 2019b; Zhao et al. 2019; Gan et al. 2020), their design procedures of fusion modules are laborious and require extensive prior knowledge from human experts.

Naturally, a question arises from such a situation: *Is there*

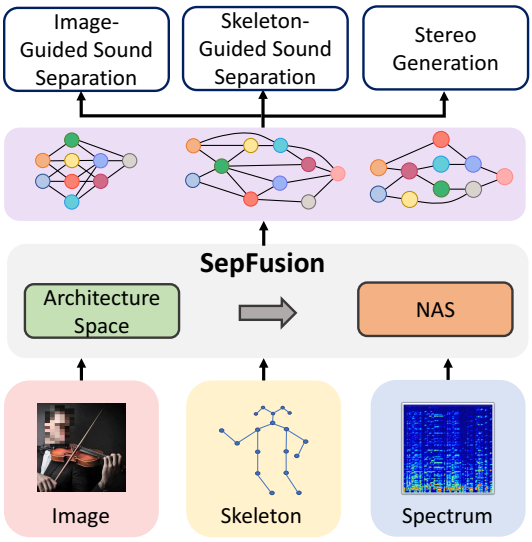

Figure 1: We propose SepFusion to smoothly construct optimal fusion structures for visual-guided audio processing tasks. We devise an attention-based architecture space for fusion structure generation and adopt Neural Architecture Search (NAS) algorithms to select the good architectures among all candidates. Our approach enjoys the advantage of flexibility, which can be deployed on different modalities and various separation scenarios.

*These authors contributed equally.
†Di Hu is the corresponding author.

any approach to easily find ideal fusion structures for the visual sound separation system?

The answer is **Yes**. In this work, we propose SepFusion, a novel framework that can smoothly produce audio-visual fusion architectures for visual sound separation. The SepFusion framework consists of two modules, namely the model generator and the evaluator. For the model generator, we design an architecture space, from which we can conveniently obtain various fusion modules based on a sampling rule. The evaluator is constructed on the Neural Architecture Search (NAS) algorithm. It can effectively pick superior architectures among the generated candidates. Compared with hand-crafted structures, this automatic process significantly saves human efforts on model design whilst achieving superior performances.

Our SepFusion framework enjoys the advantage of flexibility, which is reflected in the following aspects. (1) The framework does not have special requirements on the input modes, which means we can smoothly choose the inputs from commonly-used modalities such as image, spectrogram, skeleton, etc., according to our needs. (2) Compared with previous works (Pérez-Rúa et al. 2019; Yu et al. 2020), which aim to search for the entire network architectures, our fusion structures are relatively independent modules. Thus, our SepFusion also possesses good compatibility and can be easily plugged into existing pipelines. (3) What's more, in our architecture space, the majority of the fusion operations are primitive matrix computations, which are parameter-free. This property makes our fusion structure very lightweight, and hence relieves the computational burdens when introducing additional modules.

In addition to the advantages of flexibility and low-cost, our SepFusion also exhibits strong robustness, manifested in its versatility to various scenarios. Experimental results demonstrate that our method improves baseline methods by a large margin on the visual sound separation task. Besides, our method can be directly applied to the visually guided binaural audio generation task (Gao and Grauman 2019a; Morgado et al. 2018) as well. Furthermore, since SepFusion can simultaneously provide a series of good architectures, it can be easily benefited from the model ensemble. The experimental results show that the model ensemble not only brings extra improvements but may also be used to meet more customized requirements, such as the special demand for separating a certain instrument.

Our contributions can be summarized as followed: (1) We put forward SepFusion, a novel framework that can automatically construct lightweight audio-visual fusion modules. (2) The flexibility and robustness of our SepFusion framework are verified on visual sound separation with modalities. Extensive experiments demonstrate the effectiveness of our fusion modules. (3) We raise an effective model ensemble mechanism to fully utilize the advantages of the architecture family, which can satisfy more specific requirements whilst further boost the performances.

Related Works

Audio-Visual Learning. In recent years, audio-visual learning has attracted widespread attention with the success of deep learning. By leveraging the richer semantic information provided by two modalities, deep models tend to exhibit better performances compared with the single-modality scene. Many areas have also achieved rapid developments, such as audio-visual corresponding learning (Korbar, Tran, and Torresani 2018; Owens and Efros 2018; Arandjelovic and Zisserman 2017; Arandjelović and Zisserman 2018), cross-modality generation (Zhou et al. 2018; Oh et al. 2019; Ginosar et al. 2019; Gao and Grauman 2019a; Morgado et al. 2018; Zhou et al. 2019a,b, 2021), sound separation and localization (Zhao et al. 2018, 2019; Ephrat et al. 2018; Afouras, Chung, and Zisserman 2018; Rouditchenko et al. 2019; Gan et al. 2020; Gao, Feris, and Grauman 2018; Xu, Dai, and Lin 2019; Tian, Hu, and Xu 2021; Gao and Grauman 2021, 2019b; Hu, Nie, and Li 2019; Hu et al. 2020; Rahman, Yang, and Sigal 2021; Majumder, Al-Halah, and Grauman 2021) and so on. Our work focuses on two important tasks of the visual-guided sound processing field, namely, visual music separation and stereo sound generation (Gao and Grauman 2019a; Zhou et al. 2020b; Xu et al. 2021). Despite the success in the previous works, the effective fusion mechanisms of vision and audio features are still less explored. Zhao et al. (2019) and Gan et al. (2020) have designed specific fusion modules to enhance the audio-visual interaction. But these architectures are still rigid, making them hard to fit different tasks or modalities. Our SepFusion enjoys flexibility and can easily dig the fusion mode between different modalities.

Multi-Modality Fusion. Learning effective fusion mechanisms is one of the core challenges in multi-modality learning. By capturing the interactions between different modalities more reasonably, the deep models can accomplish the semantic compensations and hence acquire more comprehensive information. Many powerful architectures have been proposed in the video understanding field, which majorly aims at combining appearance and motion features (Simonyan and Zisserman 2014; Wang et al. 2015; Feichtenhofer, Pinz, and Zisserman 2016; Ryoo et al. 2019; Feichtenhofer et al. 2019). Both the appearance and motion modalities belong to the vision domain, while basically larger gaps may appear between cross-censor modalities. Some works try to solve the cross-sensor modality fusion problem by finding ideal feature connections in VQA (Gao et al. 2019; Yu et al. 2020) and audio-visual classification tasks (Pérez-Rúa et al. 2019). These previous works mainly focus on the high-level scenes, but our work can deal with the fine-grained dense prediction tasks on the pixel level.

Neural Architecture Search. Neural Architecture Search (NAS) aims at searching for optimal neural networks automatically in an elaborately designed search space, and its effectiveness has been verified in many computer vision fields, such as image classification (Zoph and Le 2016; Zoph et al. 2018; Xie et al. 2019; Zhou et al. 2020a), object detection (Ghiasi, Lin, and Le 2019; Xu et al. 2019) and so on. These previous works primarily target homogeneous features of similar semantics, but our work can handle more diverse features even of various sensor domains. Some works (Pérez-Rúa et al. 2019; Yu et al. 2020) also explore the application of NAS algorithms to

find structures for features from different modalities. The searching objective of these works is the entire processing system, while our SepFusion focuses on an independent fusion module, which is more compatible with existing frameworks.

Our Approach

The objective of visual-sound separation is to separate individual audio components from a mixed audio signal with the guidance of visual cues. Since the separation results are guided by visual signals, it is essential to find a reasonable information fusion mode between visual and audio modalities for the processing system. Instead of proposing a specific fusion structure as (Zhao et al. 2018, 2019; Gan et al. 2020), our work develops a more general mechanism, where the design scope is extended from a single architecture to the architecture space. Our audio-visual fusion architecture space is mainly based on primitive attention operations and can easily produce various fusion architectures with given sampling rules. After the establishment of the architecture space, we adopt a Neural Architecture Search (NAS) algorithm to select the optimal structures from the generated candidates. Compared with the hand-crafted networks, our automatic searching mechanism can handle different visual modality types (e.g., images, skeletons) and achieve stronger performances.

In this section, we will introduce how to generate new structures based on our designed architecture space. Details on the search algorithm will be presented in the next section.

Framework Overview

Following previous studies (Zhao et al. 2018; Gao and Grauman 2019b; Zhao et al. 2019), we also utilize the 'Mix-and-Separate' paradigm to construct our visual sound separation pipeline in a self-supervised manner. Given two video clips $\{V_A, V_B\}$ with corresponding audio signals $\{S_A, S_B\}$, we mix the audio components to generate a synthetic mixture audio signal $S_m = (S_A + S_B)/2$. For each video clip i ($i \in \{A, B\}$), the visual encoder extracts the visual feature f_v^i from the input frames. Meanwhile, the mixed audio signal is fed into the audio encoder to generate the audio feature f_a. Afterward, the audio feature will be fused with the visual features $\{f_v^A, f_v^B\}$, respectively, and produce audio-visual features $\{f_{av}^A, f_{av}^B\}$ responsible for separating the corresponding audio signal. Finally, the fused audio-visual features will pass through the audio decoder to produce the separation mask for each component. The pipeline is depicted in Figure 2.

Steps for Intermediate Feature Generation

The features involved in the fusion structure are termed as **nodes** which include both the input features and intermediate features. The intermediate nodes are created one by one in sequence until their amount reaches the pre-defined value. Each new node will take two previous nodes as inputs, and combine the inputs to compute its own output via a certain fusion operation. In Figure 2 (a), we provide an example of possible connections between the nodes. In the beginning, the node set only contains the monomodal audio and visual input nodes. The generation steps of each intermediate node are as follows, and the selecting action in each step is random selection.

- **Step 1.** Select two inputs from the existing node set.
- **Step 2.** Select an operation from the operation pool.
- **Step 3.** Apply the operation on the two input nodes to compute the non-activated feature.
- **Step 4.** Select one activation function from sigmoid, softmax, ReLU, and identity (no activation). Activate the feature from Step 3 to produce the new node.
- **Step 5.** Add the new node to the node set.

If not specified, we assume that the modality of visual data are frames, so that the input visual node has the shape $f_v \in \mathbb{R}^{C\times H\times W}$ (the video index is omitted for simplicity), where C, H, W refer to Channel, Height, and Width, respectively. As for the audio data, we transfer the raw signals to spectrograms via Short Time Fourier Transform (STFT). Thus, the input audio node possesses the format of $f_a \in \mathbb{R}^{C\times F\times T}$, where C, F, T stand for Channel, Frequency, and Time, respectively.

Fusion Operations

Before the demonstration of fusion operations, we will first introduce the format adjustment operation (denoted as FA), which serves as a necessary step in many fusion operations, and discuss the specific operations subsequently.

Format Adjustment The FA operation may adjust the shapes of the input nodes so that they can be used to generate the output of target formats. To obtain the output format, two principles are proposed: (1) For most cases, where both inputs contain the C dimension, the format of the generated node should be the same as one of the inputs; (2) If one of the inputs does not possess the C dimension, *i.e.* of the shape $\mathbb{R}^{H\times W\times F\times T}$, then shared dimensions of the two inputs will be eliminated and the output format is composed of the remaining dimensions. The output formats corresponding to all possible input combinations are presented in Figure 3.

The FA operation can be regarded as a 'maximize-and-replicate' procedure, that is, a feature first conducts the global max-pooling on the dimensions to be eliminated, and then obtains the target shape by replicating this maximum value. Suppose the current feature f_i is of shape $[C, M, N]$ while the target shape is $[C, J, K]$. The initial step is to reshape f_i to $[C, (M \times N)]$ and acquire the maximum value for each $M \times N$ array along the C dimension. Then we duplicate these maximum values for $J \times K$ times and reshape the feature to the new shape $[C, J, K]$. In this way, we can realize the format transformation on features, which can be formulated in Eq 1. Please note that we do not require that $[M, N]$ and $[J, K]$ must be different. In other words, the FA operation is also applicable to the cases where the target shape is the same as the original shape.

$$f_i' = \underset{[C,M,N]\mapsto[C,J,K]}{FA}(f_i) \tag{1}$$

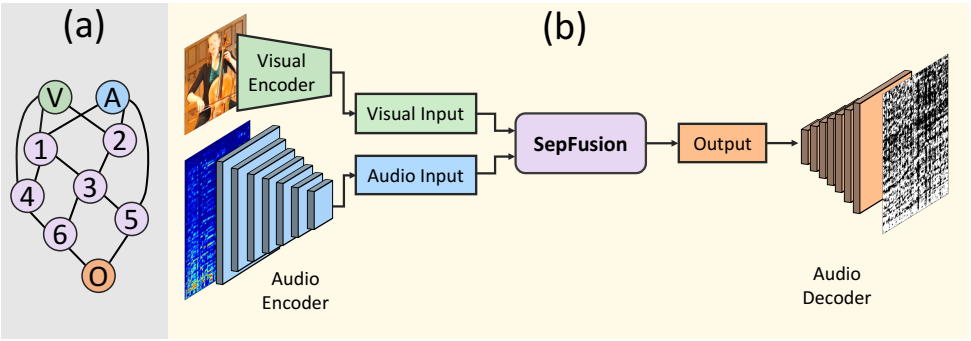

Figure 2: Sketch of one fusion module structure (a) and the illustration of the whole pipeline for the visual sound separation task (b). In the sketch, we provide an example of the connection mode for the nodes inside the fusion module. The separation process can be guided by different types of visual cues and we adopt the image modality as an example here. The audio and visual features are extracted by the corresponding encoders and then get fused in the SepFusion module. Finally, the fused audio-visual feature will pass through the audio decoder to predict the results.

Specific Operations The specific fusion operations will be discussed as follows. As shown in Figure 3, the constitution of the *operation pool* during the generation will change with different input combinations. The operations are *element-wise Addition (EmADD), element-wise Multiplication (EmMUL), Multiplication (MUL), Concatenation (Concat), and Skip-connection (SkC).* We can see that most operations are based on matrix operations, which makes the fusion module very lightweight and saves the computational budgets.

EmADD: The element-wise addition requires that the two inputs should share the same formats. Therefore, we need to make sure the shapes of the two inputs are consistent before conducting the operation. Suppose the two input nodes are $f_X \in \mathbb{R}^{C\times M\times N}$ and $f_Y \in \mathbb{R}^{C\times J\times K}$ while the output node is f_Z. For the situation where $M \neq J$ and $N \neq K$, which corresponds to the combination of $\{[C,H,W]\&[C,F,T]\}$ in Figure 3, we need to check the required format of the output f_Z. If the format of f_Z is the same as that of f_X, we need to adjust the shape of f_Y via the FA operation, and vice versa. We depict this operation in Eq 2.

$$\begin{aligned} f_Z &= EmADD(f_X, f_Y) \\ &= f_X + \underset{[C,J,K]\mapsto[C,M,N]}{FA}(f_Y), \quad f_Z \in \mathbb{R}^{C\times M\times N} \end{aligned} \tag{2}$$

When the two inputs already share the same format, that is, the combination of $\{[C,H,W]\&[C,H,W]\}$ or $\{[C,F,T]\&[C,F,T]\}$, we still choose to apply the FA operation on one of the inputs while keeping another one unchanged. In this way, the FA operation does not serve as the shape transformation method but can be regarded as an attention mechanism. This situation is formulated in Eq 3.

$$\begin{aligned} f_Z &= EmADD(f_X, f_Y) \\ &= f_X + \underset{[C,M,N]\mapsto[C,M,N]}{FA}(f_Y), \quad f_Z \in \mathbb{R}^{C\times M\times N} \end{aligned} \tag{3}$$

EmMUL: The element-wise multiplication operation is very similar to the EmADD operation, except that the operator changes from addition to multiplication. We adopt the same rules to adjust the input nodes before applying the computation.

MUL: Different from the EmMUL operation, which requires the two inputs have the same formats, the MUL operation will eliminate all the common dimensions of the two input nodes, and only retain the distinct dimensions. Therefore, this operation is forbidden when the formats of two inputs are exactly the same, which is also reflected in Figure 3. In practice, the Einstein Summation Convention is utilized to accomplish dimension compression. Suppose the two inputs are $f_X \in \mathbb{R}^{C\times M\times N}$ and $f_Y \in \mathbb{R}^{M\times N\times J\times K}$, where the common dimensions are M and N. By compressing the common dimensions and combine the distinct ones, we can obtain the output format $f_Z \in \mathbb{R}^{C\times J\times K}$. Each element of feature f_Z can be computed as follows:

$$f_Z^{c,j,k} = \sum_m \sum_n f_X^{c,m,n} \times f_Y^{m,n,j,k} \tag{4}$$

Concat: This operation concatenates the inputs along the Channel dimension. Thus, the existence of C dimension is necessary for both input nodes. Naturally, the input formats should first be adjusted to match each other. When the input formats are different, we adopt a similar format adjustment method as in *EmADD* and *EmMUL* to realize the matching. However, when the input formats are already the same, such as the combination of $\{[C,H,W]\&[C,H,W]\}$, the adjustment will be skipped and the two input nodes will be directly concatenated. Finally, the concatenated feature passes through a 1×1 convolution layer to recover the channel number from $2C$ to C and generate the output node.

SkC: The skip-connection operation adapts to a relatively special case, when the two selected input nodes happen to be the same node and the Channel dimension exists. We feed the input node to a 1×1 convolution layer then add the convolved feature and the original input to produce the output.

Output of Fusion Module

After the generation of all the intermediate nodes, we can build the final output feature for the entire fusion module.

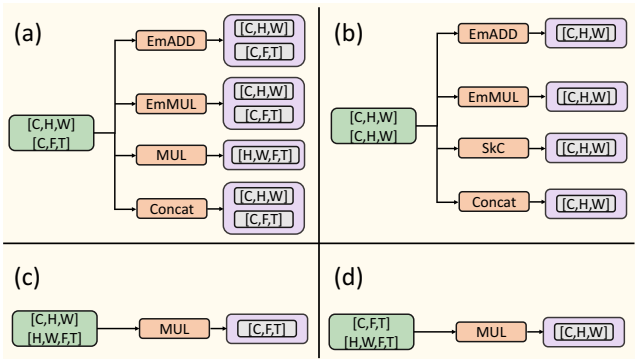

Figure 3: Illustration of all permitted input node combinations (green rectangles), as well as the operation pools (orange rectangles) and the corresponding formats of the output node (grey rectangles). The modality of visual data are frames while the audio data are spectrograms. We can see that the operation pools will change with different input combinations. Please note that the combination of $\{[C, F, T], [C, F, T]\}$ is omitted in the figure since it is the same as the $\{[C, H, W], [C, H, W]\}$ group.(only need to replace all $[C, H, W]$ to $[C, F, T]$ in (b)). Best view in color.

The nodes that never serve as inputs to other nodes should appear in the output set to make sure that no generated nodes are useless, unless they do not contain the $Channel$ dimension. Please note that the shape of nodes inside the output set may be different, so that we need to convert the features into the specific output format. In our task, the subsequent processing is operated on the audio domain. Thus, the nodes with visual shapes (*e.g.* $\mathbb{R}^{C\times H\times W}$) are transformed to be consistent with the audio format via the FA operation. Finally, we concatenate these nodes of pure audio semantics and convolve them to obtain the final output of the fusion module. Figure 4 shows an example of the complete fusion module structure.

Search Algorithm

In the previous section, we illustrate the steps to randomly generate the audio-visual fusion structures. To obtain the ideal fusion modules, we adopt an evolution-based search algorithm that evaluates and selects the optimal architectures among all the possible candidate networks.

In the searching process, the initial fusion architectures are randomly generated. After that, the new architectures will be constructed from the mutation of existing structures in the $population$ set, which are denoted as $parent$ models. All the searched structures and their performances are stored in the $history$ set.

For every evolution cycle, we select the network with the highest performance from the current $population$ set as the $parent$ model. After the mutation manipulation, the $parent$ model generates a new $child$ architecture. We train and evaluate the new $child$ to obtain the performance, and then add it to both the $history$ and $population$ set. Meanwhile, the oldest model (not necessarily the worst) is removed from the $population$ set. Thus, the $population$ is dynamically updated with the evolutionary process, instead of keeping constant. In this way, the candidate networks may enjoy a higher diversity, rather than dominated by a certain constant parent. The details of the algorithm are shown in the supplementary.

Following previous evolutionary-based NAS methods (Real et al. 2019; Liu et al. 2017; Real et al. 2017), we utilize two mutation methods that we call *input mutation* and *operation mutation*. Both mutation methods will only change a single node in the entire fusion module. The *input mutation* will replace one of the inputs with another one, which can be selected from the nodes previous to the current mutation node, while the *operation mutation* keeps the inputs unchanged and modify the fusion operation. The illustration of the two mutation methods is shown in Figure 5.

Model Configuration

In this section, we provide details about the configuration of the models in the visual sound separation task.

Vision Network

We use dilated ResNet-18 (He et al. 2016) network to extract the frame features. The final global pooling and fully-connected layers are removed from the network so that the features from the 4^{th} ResNet block are regarded as the final output. The output shape is $C_v \times (H_0/32) \times (W_0/32)$, where H_0 and W_0 denote the height and width of input frames and C_v stands for channel dimension.

Audio Network

Following previous works (Zhao et al. 2018, 2019; Gao and Grauman 2019b,a), we adopt a U-Net (Ronneberger, Fischer, and Brox 2015) style architecture as the audio network. The U-Net consists of the same number of downsample layers and upsample layers with skip connections between feature maps of the same scale. The input to the audio network is a 2D Time-Frequency magnitude spectrogram of mixed audio signal computed by STFT. The audio features after the downsample layers together with the visual features are fed into the fusion module to complete the cross-modality interaction. Afterward, the fused features will pass through the upsample layers to predict the separation masks for the audio components.

The entire framework is optimized with the pixel-level sigmoid binary cross-entropy loss, where the ground truth mask for each component is obtained by checking whether the corresponding audio signal is dominant in the mixed audio at each T-F unit.

Experiments

In this section, we first introduce the datasets and evaluation metrics. The searching and training details will be provided in the supplementary. Then we compare the performance of SepFusion with the baseline methods and show the ablation studies. Finally, we propose some interesting applications of the searched models, which will further promote the performances or provide suitable networks for specific instrument classes.

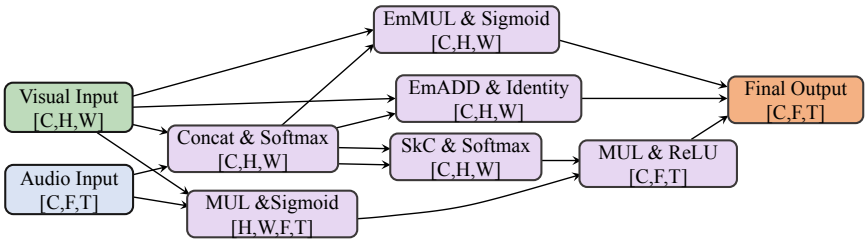

Figure 4: An example of the randomly generated fusion module, which consists of 6 intermediate nodes. The visual input is from image modality with format $[C, H, W]$ while the audio input is from Time-Frequency spectrogram with format $[C, F, T]$. We mark the operation, activation as well as the output format on the intermediate nodes (the violet rectangles).

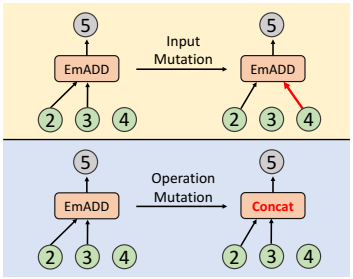

Figure 5: Illustration of the two mutation methods to modify a node. The green circles represent the inputs to a node while the outputs are shown in gray circles. The numbers serve as the node index.

Datasets and Evaluation Metrics

We validate the effects of our fusion module on the MUSIC dataset (Zhao et al. 2019), which contains 21 classes of instruments. The dataset consists of untrimmed videos crawled from the YouTube website and hence can be regarded as separation in the wild. We adopt the open-source mir_eval library (Raffel et al. 2014) to compute the following metrics: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). The units are dB.

Performance Comparison

We compare with existing baseline methods for visual sound separation and summarize the results in Table 1. Sound-of-Pixels (Zhao et al. 2018) executes the audio-visual fusion at the end of U-Net. Simple-concat is a self-implemented baseline which moves the audio-visual fusion operation to the middle of the U-Net while the remaining parts are kept the same as Sound-of-Pixels. The fusion operation is concatenation, as the name suggests. Co-separation (Gao and Grauman 2019b) incorporates an object-level co-separation loss for the separation framework.

We illustrate the comparison between SepFusion and the baseline methods in Table 1, where we can observe that our approach consistently exceeds all baselines on different metrics. Specifically, our fusion module surpasses the most competitive baseline by 1.0 dB on the SDR score. We also argue that our SepFusion is mainly based on matrix operations so basically it only brings limited extra parameters. In other words, we achieve impressive performance improvements at a low cost of parameter burdens.

Method	SDR↑	SIR↑	SAR↑
Simple-Concat	7.30	13.22	10.65
Sound-of-Pixels (2018)	7.57	14.20	11.48
Co-Separation (2019b)	7.76	12.93	10.89
SepFusion (Ours)	**8.76**	**16.65**	**11.81**

Table 1: Separation results on the MUSIC test set. Higher is better for all metrics.

Ablation Study

Verification on other visual modality. In addition to the image modality, we also examine the compatibility of our SepFusion with skeletons as the visual information, as illustrated in Table 2. Skeleton-based music separation was first proposed by Gan *et al.* in (Gan et al. 2020) together with a devised Visual-Audio Fusion Module (denoted as VAFM). We make a comparison with both the simple baseline of concatenation (denoted as Simple-Concat) and the hand-crafted VAFM. The implementation details are provided in the supplementary material. The results exhibit that our fusion module is more effective than baselines. Besides, our SepFusion also possesses a unique advantage of flexibility, as it can provide a series of good fusion structures at the same time while VAFM only represents single architecture. The flexibility shows potential capabilities for a broader range of applications, which will be discussed in the following part.

Method	SDR↑	SIR↑	SAR↑
Simple-Concat	9.02	16.70	11.97
VAFM (Gan et al. 2020)	9.28	17.41	11.44
SepFusion (Ours)	**9.71**	**18.05**	**12.20**

Table 2: Separation results on the MUSIC test set with skeleton modality as visual cues.

Verification on other separation scenario. In addition to the normal separation task, we also examine the SepFusion performance on another separation scenario, i.e., the stereo generation task (Gao and Grauman 2019a). The stereo generation task shares similar network architectures with the

separation task but the outputs are binaural masks. We conduct experiments on the Fair-Play dataset and the results are listed in Table 3. Other implementation details can be found in the supplementary material. We observe that the models with our fusion modules can shorten both the spectrogram and envelope distances, which exhibit the effectiveness of our SepFusion framework. With the fusion module, the models can better exploit the spatial layout and appearance information provided by the visual cues, which may lead to the generation of higher quality binaural sounds.

Method	$STFT_D \downarrow$	$ENV_D \downarrow$
Mono2Binural (2019a)	0.959	0.141
SepFusion (Ours)	**0.927**	**0.137**

Table 3: Stereo generation results on the test split of Fair-Play dataset. Lower is better for all metrics.

Ensemble of Searched Structures

As discussed, our SepFusion enjoys the advantage of flexibility as it can produce a basket of relatively good structures simultaneously. In this part, we will present two choices on how to exploit the strengthens of the model library. First, different architectures can serve as complements to each other and promote overall performance by compensating for the shortcoming of every single model. Second, the diversified structures may respond to more delicate and customized demands, such as the prominent enhancement of certain instruments. These bonus benefits further boost the competitiveness of our approach.

By combining the advantages of different fusion structures, we can surpass the upper limit of a single model and obtain more impressive outcomes. This capacity is validated on both sound separation and stereo generation tasks, as depicted in Table 4. For the sound separation task, we assemble the masks predicted by every single model and take the maximum value at each pixel to generate the integrated mask. Then the integrated mask will be binarized using the same threshold as the single model to generate the final output. For the stereo side, we conduct mean operation on binaural sounds generated by each model to reduce the fluctuations in individually predicted waveforms. The results in Table 4 demonstrate that the improvements are robust and consistent on different tasks, which also lead to a new solution for performance promotion.

Naturally, different architectures should have distinctive characteristics, which allows the architecture pool to deal with more specific requests, such as target separation at one certain instrument. We observe that the overall best-performing model does not necessarily exhibit superior ability on every single instrument. Accordingly, a relatively sub-optimal model may still possess an advantage in a certain category. We visualize our observation in Figure 6, which serves as a good example to illustrate the preferences of different structures on distinct instruments. We can observe that the overall best model does not necessarily perform best in every category, while the less optimal model may still show superior ability on some instrument. The relevant spectrogram visualizations are provided in the supplementary.

Separation	SDR↑	SIR↑	SAR↑
Single Best	8.76	16.65	11.81
Ensemble	**9.20**	**17.14**	**12.04**
Stereo	$STFT_D \downarrow$	$ENV_D \downarrow$	
Single Best	0.927	0.137	
Ensemble	**0.895**	**0.136**	

Table 4: Comparisons between results predicted by the single best model and the model ensemble on both sound separation and stereo generation tasks.

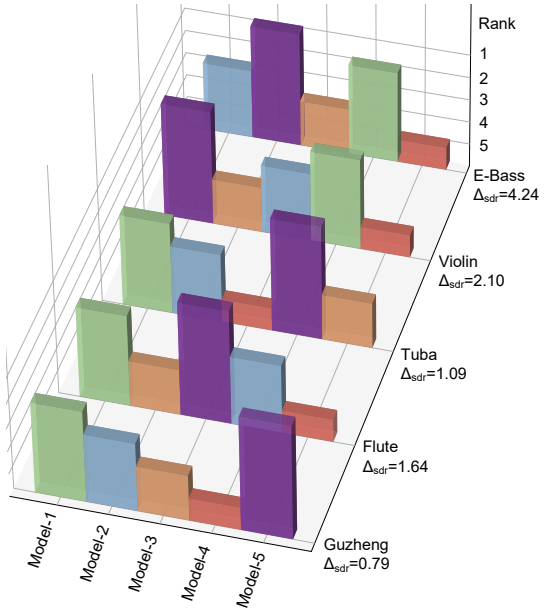

Figure 6: Performance visualization of different architectures on distinct instruments. The model index represents the overall performance rank. The Rank-axis denotes the SDR ranking for separating a certain instrument and a higher bar indicates better performance. Δ_{sdr} refers to the SDR differences between the best and worst models.

Conclusion

We propose SepFusion, a framework that can easily produce optimal audio-visual fusion structures. Experimental results demonstrate that our SepFusion surpasses baseline methods on various scenarios and also exhibits advantages when adopting visual cues of different modalities, which proves the robustness of our approach. Besides, our SepFusion enjoys an important advantage that can provide a series of relatively good architectures instead of a single structure. This property may bring interesting applications, such as model ensemble and customized architecture, which to our knowledge have never been realized by previous works. We wish this work could enlighten more discussions and explorations about the cross-modality feature fusion mechanism.

Acknowledgements

This work was supported in part by the Research Funds of Renmin University of China (NO. 21XNLG17 and 2021030200) and the 2021 Tencent AI Lab Rhino-Bird Focused Research Program (NO. JR202141). This work was also supported in part by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). Wanli Ouyang was supported by the Australian Research Council Grant DP200103223, FT210100228, and Australian Medical Research Future Fund MRFAI000085, CRC-P "ARIA - Bionic Visual-Spatial Prosthesis for the Blind" and "Smart Material Recovery Facility (SMRF) - Curby Soft Plastics", and SenseTime.

References

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. The Conversation: Deep Audio-Visual Speech Enhancement. *Proceedings of the Interspeech*.

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, 609–617.

Arandjelović, R.; and Zisserman, A. 2018. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 435–451.

Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W. T.; and Rubinstein, M. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, 6202–6211.

Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1933–1941.

Gan, C.; Huang, D.; Zhao, H.; Tenenbaum, J. B.; and Torralba, A. 2020. Music Gesture for Visual Sound Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10478–10487.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6639–6648.

Gao, R.; Feris, R.; and Grauman, K. 2018. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 35–53.

Gao, R.; and Grauman, K. 2019a. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 324–333.

Gao, R.; and Grauman, K. 2019b. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision*, 3879–3888.

Gao, R.; and Grauman, K. 2021. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2019. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7036–7045.

Ginosar, S.; Bar, A.; Kohavi, G.; Chan, C.; Owens, A.; and Malik, J. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3497–3506.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, D.; Nie, F.; and Li, X. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9248–9257.

Hu, D.; Qian, R.; Jiang, M.; Tan, X.; Wen, S.; Ding, E.; Lin, W.; and Dou, D. 2020. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33.

Korbar, B.; Tran, D.; and Torresani, L. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, 7763–7774.

Liu, H.; Simonyan, K.; Vinyals, O.; Fernando, C.; and Kavukcuoglu, K. 2017. Hierarchical representations for efficient architecture search. *arXiv preprint arXiv:1711.00436*.

Majumder, S.; Al-Halah, Z.; and Grauman, K. 2021. Move2Hear: Active Audio-Visual Source Separation. *arXiv preprint arXiv:2105.07142*.

Morgado, P.; Nvasconcelos, N.; Langlois, T.; and Wang, O. 2018. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*.

Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7539–7548.

Owens, A.; and Efros, A. A. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 631–648.

Pérez-Rúa, J.-M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 6966–6975.

Raffel, C.; McFee, B.; Humphrey, E. J.; Salamon, J.; Nieto, O.; Liang, D.; Ellis, D. P.; and Raffel, C. C. 2014. mir_eval:

A transparent implementation of common MIR metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer.

Rahman, T.; Yang, M.; and Sigal, L. 2021. TriBERT: Full-body Human-centric Audio-visual Representation Learning for Visual Sound Separation. *arXiv preprint arXiv:2110.13412*.

Real, E.; Aggarwal, A.; Huang, Y.; and Le, Q. V. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, 4780–4789.

Real, E.; Moore, S.; Selle, A.; Saxena, S.; Suematsu, Y. L.; Tan, J.; Le, Q. V.; and Kurakin, A. 2017. Large-scale evolution of image classifiers. In *International Conference on Machine Learning*, 2902–2911. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Rouditchenko, A.; Zhao, H.; Gan, C.; McDermott, J.; and Torralba, A. 2019. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2357–2361. IEEE.

Ryoo, M. S.; Piergiovanni, A.; Tan, M.; and Angelova, A. 2019. Assemblenet: Searching for multi-stream neural connectivity in video architectures. *arXiv preprint arXiv:1905.13209*.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, 568–576.

Tian, Y.; Hu, D.; and Xu, C. 2021. Cyclic Co-Learning of Sounding Object Visual Grounding and Sound Separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, L.; Xiong, Y.; Wang, Z.; and Qiao, Y. 2015. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*.

Xie, S.; Kirillov, A.; Girshick, R.; and He, K. 2019. Exploring randomly wired neural networks for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 1284–1293.

Xu, H.; Yao, L.; Zhang, W.; Liang, X.; and Li, Z. 2019. Auto-fpn: Automatic network architecture adaptation for object detection beyond classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 6649–6658.

Xu, X.; Dai, B.; and Lin, D. 2019. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Xu, X.; Zhou, H.; Liu, Z.; Dai, B.; Wang, X.; and Lin, D. 2021. Visually Informed Binaural Audio Generation without Binaural Audios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, Z.; Cui, Y.; Yu, J.; Wang, M.; Tao, D.; and Tian, Q. 2020. Deep Multimodal Neural Architecture Search. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3743–3752.

Zhao, H.; Gan, C.; Ma, W.-C.; and Torralba, A. 2019. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, 1735–1744.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*.

Zhou, D.; Zhou, X.; Zhang, W.; Loy, C. C.; Yi, S.; Zhang, X.; and Ouyang, W. 2020a. Econas: Finding proxies for economical neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11396–11404.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019a. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhou, H.; Liu, Z.; Xu, X.; Luo, P.; and Wang, X. 2019b. Vision-infused deep audio inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, H.; Xu, X.; Lin, D.; Wang, X.; and Liu, Z. 2020b. Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; and Berg, T. L. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zoph, B.; and Le, Q. V. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Zoph, B.; Vasudevan, V.; Shlens, J.; and Le, Q. V. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710.