SeCo: Separating Unknown Musical Visual Sounds with Consistency Guidance Supplementary Material

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

A. Architecture of the Vision Analysis Network

In Table. A1, we show the architecture details of the vision analysis network. The kernel shape is denoted as $\{T \times {S_z}^2, C\}$ for temporal, spatial and channel dimensions, respectively. We do not perform the downsampling operation on the temporal dimension in the entire structure so that the temporal stride is fixed to 1. The parameter α in the bottleneck structure represents the channel reduction ratio and is set to $1/4$.

Structure	Information			
Bottleneck	$\{3\times1^2, \alpha C\}$ Conv3d, BN, ReLU $\{1\times3^2, \alpha C\}$ Conv3d, BN, ReLU $\{1\times1^2, C\}$ Conv3d, BN			
Vision Network	Stage	Operator	Stack	Stride
	Input	-	-	-
	Conv1	$\{5\times7^2, 8\}$	-	2
	Pool1	1×3^2, MaxPool3d	-	2
	Res1	Bottleneck, C=32	3	1
	Res2	Bottleneck, C=64	4	2
	Res3	Bottleneck, C=128	6	2
	Res4	Bottleneck, C=256	3	2
	Pool2	Global AvgPool3d	-	-

Table A1. Implementation details of the vision analysis network. 'Stack' refers to the number of stacked bottleneck blocks in each residual stage and 'Stride' to the stride value along the spatial dimensions for the conv3d or pooling operations.

B. Algorithm for Online Matching Strategy

In Alg. A1, we illustrate the optimization process of the online matching strategy.

Algorithm A1

Require: $\mathcal{P}$: the testing dataset
Require: θ_0: Pre-trained model parameters
Require: $\mathcal{T}$, β: hyper-parameters for online matching
for p_k in $\mathcal{P}$ **do** ▷ p_k refers to single test pair
  for τ = 1 to $\mathcal{T}$ **do**
    Compute the gradient $\bigtriangledown_\theta L_{cs}$ using $\theta_{\tau-1}$ and p_k
    Update the parameters: $\theta_\tau = \theta_{\tau-1} - \beta \bigtriangledown_\theta L_{cs}$
  end for
  Compute the refined separation masks: $M_k = Model(p_k|\theta_\mathcal{T})$
  Reset model parameters to θ_0
end for
return Refined separation masks for all pairs in $\mathcal{P}$

C. Data Processing Details of SeCo

During training, we randomly segment a 6-second video clip from the dataset. The audio signal is re-sampled to 11kHz. The audio signals are transformed to spectrogram by STFT with window size 512 and hop length 256. Thus, we obtain a Time-Frequency audio representation of shape 512×256. The spectrograms are further re-sampled on a log-frequency scale to produce 256×256 T-F representations as the U-Net input.

For the visual side, we take 24 frames for every video clip, whose resolution is 128×128. During training, the frames are resized to 160×160 and a region of 128×128 is randomly cropped as input. Other augmentation operations include horizontally flipping at 0.5 probability and normalization. At the inference stage, the randomly cropping operation is replaced with central cropping and the horizontal flip is skipped to reduce the effects of random factors.

During training, two video clips from different categories in the training split are randomly selected and mixed. At the inference stage, we randomly pick up 300 pairs of video clips from the testing split to construct the testing dataset.

Split	Test Class
Split-1	banjo, electric bass, guzheng, saxophone, trumpet
Split-2	bagpipe, drum, flute, piano, saxophone
Split-3	congas, guzheng, saxophone, trumpet, ukulele

Table A2. Category distribution of different splits. We list the 5 categories for testing while the remaining 16 categories serve the training set.

D. Implementation Details on the Image and Skeleton Modality

For the image modality, we use a ResNet-18 [3] network to extract the visual features after the 4^{th} ResNet stage. The convolution operations are replaced with the dilated convolution with the dilation set to 1, as in [5]. Finally, a global MaxPooling operation is applied to generate the visual feature vector. In this way, the features are only convolved on the spatial dimensions while the temporal representations are not learned.

For the skeleton modality, we use the pretrained OpenPose toolbox [1] instead of the manual annotations to estimate the 2D coordinates and the confidence scores for keypoints of the players in the video clips. Specifically, we obtain 18 joints for the human body and 21 joints for each hand. We select the Spatial-Temporal Graph CNN (ST-GCN) [4] as the vision analysis network to effectively extract features from the skeleton inputs. We follow [1] to build the edges between keypoints and form the spatial graph. The ST-GCN is composed of 11 layers with residual connections. After passing through the ST-GCN backbone, a global average-pooling operation is exerted on the feature map to produce a visual feature vector. For fair comparison of different modalities, we only adopt the skeleton modality as the visual cues instead of combining the skeleton and appearance information, which is different from [2].

E. Category Distribution of Splits

In Table. A2, we list the categories in the testing set for different splits. The other 16 categories will serve as the training set.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[2] 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.

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

[4] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018.

[5] 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.