Visually Guided Sound Source Separation using Cascaded Opponent Filter Network

Lingyu Zhu[0000−0001−6707−6665] and Esa Rahtu[0000−0001−8767−0864]

Tampere University, Tampere, Finland
lingyu.zhu@tuni.fi, esa.rahtu@tuni.fi

A Supplementary Material

A.1 Sound Source Separation for Instrument Combinations

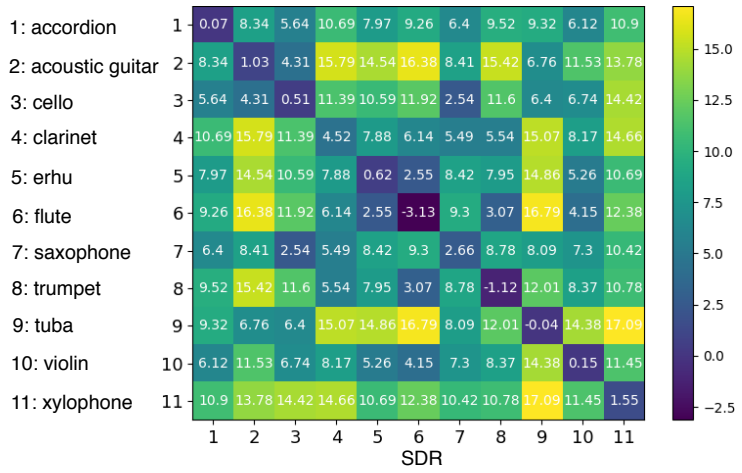

Fig. 1. The sound source separation performance for different mixtures of instruments in MUSIC dataset. The results are shown in terms of SDR. The diagonals and off-diagonals represent the results of separating instrumental combinations of same and different categories respectively. The higher value of SDR represents the better performance of sound source separation, e.g. *acoustic guitar* and *flute* (2 and 6), *flute* and *tuba* (6 and 9), *tuba* and *xylophone* (9 and 11). The SDR values on the diagonals clearly indicate the hardest case of separating sounds between two instruments from the same category, e.g. *flute* (6 and 6).

In this section, we present sound source separation performance for different instrument mixtures using MUSIC dataset. Fig. 1 illustrates the results in terms of SDR in a matrix form. The diagonals represent the results of separating instruments of same categories (e.g. two guitars), and the off-diagonals are combinations from different categories (e.g. guitar and violin). The higher

value of SDR represents the better performance of sound source separation, e.g. *acoustic guitar* and *flute* (2 and 6), *flute* and *tuba* (6 and 9), *tuba* and *xylophone* (9 and 11). The SDR values on the diagonals clearly indicate that separating sounds between two instruments from the same category is the hardest case. In particular, separating the mixture of two flutes is challenging. One reason might be the small amount of motion related to playing flute.

A.2 Sound Source Localization Examples

We visualize more examples of localized sounding sources by our proposed Sound Source Location Masking (SSLM) network in comparison with baseline methods of SoP [1], SoM [2], and MP-Net [3] on MUSIC, A-MUSIC and A-NATURAL datasets in Fig. 4, Fig. 5, and Fig. 6 respectively.

A.3 Datasets

We evaluate the proposed approaches using Multimodal Sources of Instrument Combinations (MUSIC) [1] dataset, and two sub-sets of AudioSet [4]: A-MUSIC and A-NATURAL.

MUSIC The MUSIC dataset is relatively small high quality dataset of musical instruments. It contains 714 untrimmed YouTube videos which span 11 instrumental categories, namely *accordion*, *acoustic guitar*, *cello*, *clarinet*, *erhu*, *flute*, *saxophone*, *trumpet*, *tuba*, *violin*, and *xylophone*. For all the reported experiments, we randomly split the dataset into 400 training videos, 100 validation videos, and 130 test videos.

A-MUSIC and A-NATURAL A-MUSIC dataset is a trimmed musical instrument dataset from AudioSet. It has around 25k videos spanning 10 instrumental categories: *accordion*, *bagpipe*, *cello*, *flute*, *piano*, *pizzicato*, *saxophone*, *trumpet*, *ukulele*, and *zither*. A-NATURAL dataset is a trimmed natural sound dataset from AudioSet. It contains around 10k videos which cover 10 categories of natural sounds, namely *baby crying*, *chainsaw*, *dog*, *drum*, *firework*, *helicopter*, *printer*, *rail*, *snoring*, and *water*. We split both the A-MUSIC and A-NATURAL dataset samples to 80%, 10%, and 10% as train, validation and test set.

A.4 Implementation Details

Overall Architecture We illustrate the overall architecture of the COF model in the case of two stages on two sound sources in Fig. 2. The vision networks of the COF model at different stages change accordingly to the vision network options discussed in Sec. 3.2 of main paper.

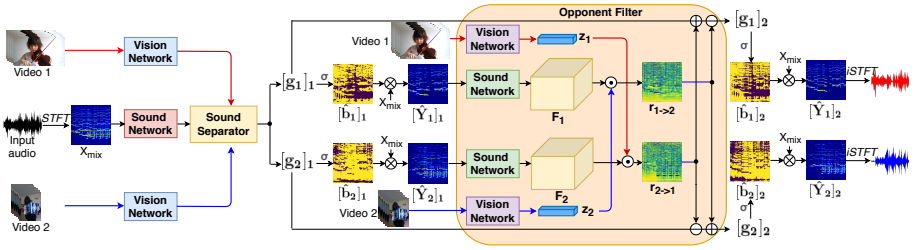

Fig. 2. The overall architecture of the COF model in the case of two stages on two sound sources. In the first stage, visual representations (vision network) and sound features (sound network) are passed to the sound separator that produces the spectrum masks g for each source. g are then further provided as inputs for the upcoming stage. Stage two refines the separation result using visual representation z_2 to identify components $r_{1->2}$ from the source 1 that should belong to the source 2. The spectrum masks g are updated accordingly by subtracting from $[g_1]_1$ and adding to $[g_2]_1$. Similar operation is done for the source 2. Finally, we obtain the separated audios by applying inverse STFT to the output component spectrograms $[\hat{Y}_1]_2$ and $[\hat{Y}_2]_2$, which are converted from $[g_1]_2$ and $[g_2]_2$ (Eq. (2) in main paper). Note that the vision networks at different stages can change accordingly to the vision network options discussed in Sec. 3.2 of main paper.

Vision Network We extract video frames at 8fps and adopt frame augmentation by random scaling, random horizontal flipping, and random cropping (224×224) during training for all datasets. We apply a dilated 2D ResNet18 [5] with *dilation*=2 to obtain representations of C2D-RGB and C2D-DYN. For a single input RGB image or dynamic image of size $3 \times 16H \times 16W$, we truncate the ResNet18 after *stride*=16 and achieve the visual feature of size $K \times H \times W$ by performing a 3×3 convolution with output channels of K=16 on the top. The C3D models utilize 3D version of ResNet18 on T=48 frames. With the *stride*=16 on spatial dimension and *stride=8* on the temporal dimension, we yield the C3D-RGB and C3D-FLO representations of size $T^{'} \times K \times H \times W$, where $T^{'}$=6 and $H = W = 14$.

Mutual Attention Module The Mutual Attention (MA) module is proposed to fuse the appearance and motion information. In the MA module, we obtain the spatial attention map by projecting the appearance features from C2D-RGB to a single-channel feature map with a 1×1 convolution and a sigmoid operation. The MA module enhances the sound source relevant motions by multiplying the C3D features with the spatial attention map. The appearance-weighted features are added back to the original C3D features in order to keep C3D features as the principle cue in case the C2D-RGB fails to localize the sound source. We obtain C3D feature attention by adding a sigmoid function on top of the final enhanced C3D features. The multiplication between the C3D feature attention and the time-inflated appearance features are added back to the C2D-RGB appearance features. Within this process, for the predicted regions of interest from C2D-

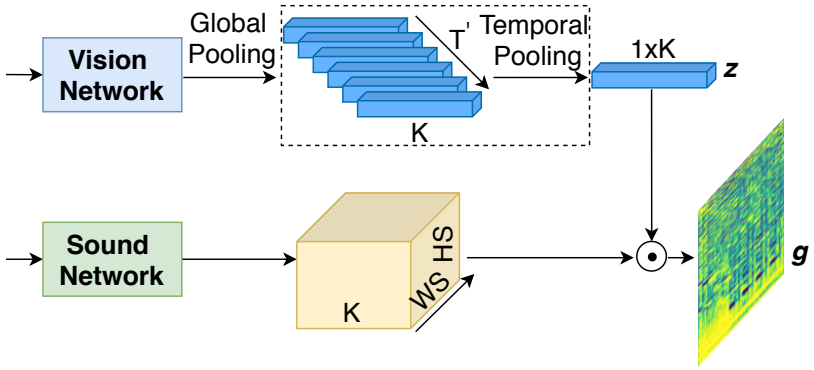

Fig. 3. The architecture of sound separator. Sound separator combines visual representations **z** with sound network feature maps using a linear combination to predict the spectrum mask g. g is then provided for the upcoming stage as input.

RGB, the appearance that has no motions will be eliminated. Finally, we receive the mutual attentive features of dimension $T^{'} \times K \times H \times W$ from the two-stream structures.

Sound Network We adopt the U-Net [6] with 7 layers of 2D CNN and output channels of *K=16* as the architecture of Sound Network. The input audio signals are represented as spectrograms, which are obtained from the audio stream using Short-time Fourier transform (STFT). To obtain the final separated audio signals, the inverse STFT is applied to the component spectrograms.

Sound Separator We depict the architecture of the sound separator (Eq. (1) in main paper) in Fig. 3. The sound separator combines the visual representations **z** with the sound network output using a linear combination to produce the spectrum mask g. Spectrum masks g for all the sources are then provided for the upcoming stage as inputs.

Sound Source Location Masking Network The SSLM network is implemented using a dilated residual network (DRN) [7] pre-trained on ImageNet [8], with three up-projection blocks [9] followed by a 3×3 convolution layer. The SSLM is trained together with the overall model in a self-supervised manner. Firstly, we train the plain COF and freeze the model parameters. Secondly, we add the "SSLM" to the COF model as shown in Fig. 5a. The input video frames are first passed through the SSLM component which outputs a weighted location mask [0,1] having same spatial size as the input frame. The input video frames are multiplied element-wise with the mask, and the result is passed to the COF model. The "SSLM" parameters are optimized to identify a minimum set of input pixels, for which the subsequent COF network output is almost identical to

the COF output without "SSLM". Finally, the "SSLM" and COF models are fine-tuned jointly.

Optimization Our implementation is built on Pytorch. The network is trained with a batch size of 10 for 4,000 iterations. We use stochastic gradient descent (SGD) with momentum 0.9 and weight decay 1e-4 to train our Cascaded Opponent Filter (COF) network and Adam optimizer to train the Sound Source Location Masking (SSLM) network. The vision networks of COF and the SSLM, pre-trained on ImageNet [8], use a learning rate of 1e-4, while the rest of modules which are trained from scratch use a learning rate of 1e-3. We decrease the learning rate from its initial value by a factor of 10 every 1,600 iterations.

References

1. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 570–586
2. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 1735–1744
3. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 882–891
4. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2017) 776–780
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
7. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 472–480
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255
9. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV), IEEE (2016) 239–248

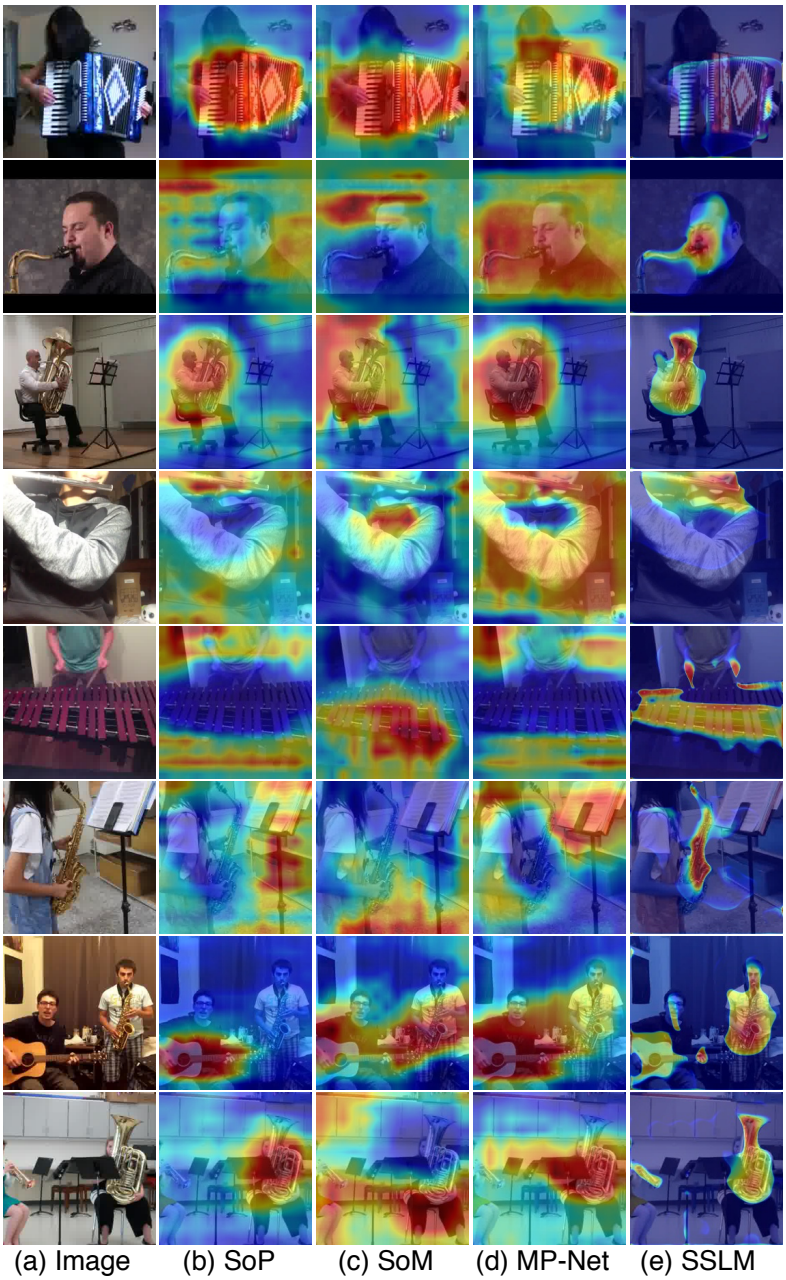

Fig. 4. Visualizing sound source location of our proposed SSLM network in comparison with baseline methods SoP, SoM, and MP-Net on MUSIC dataset.

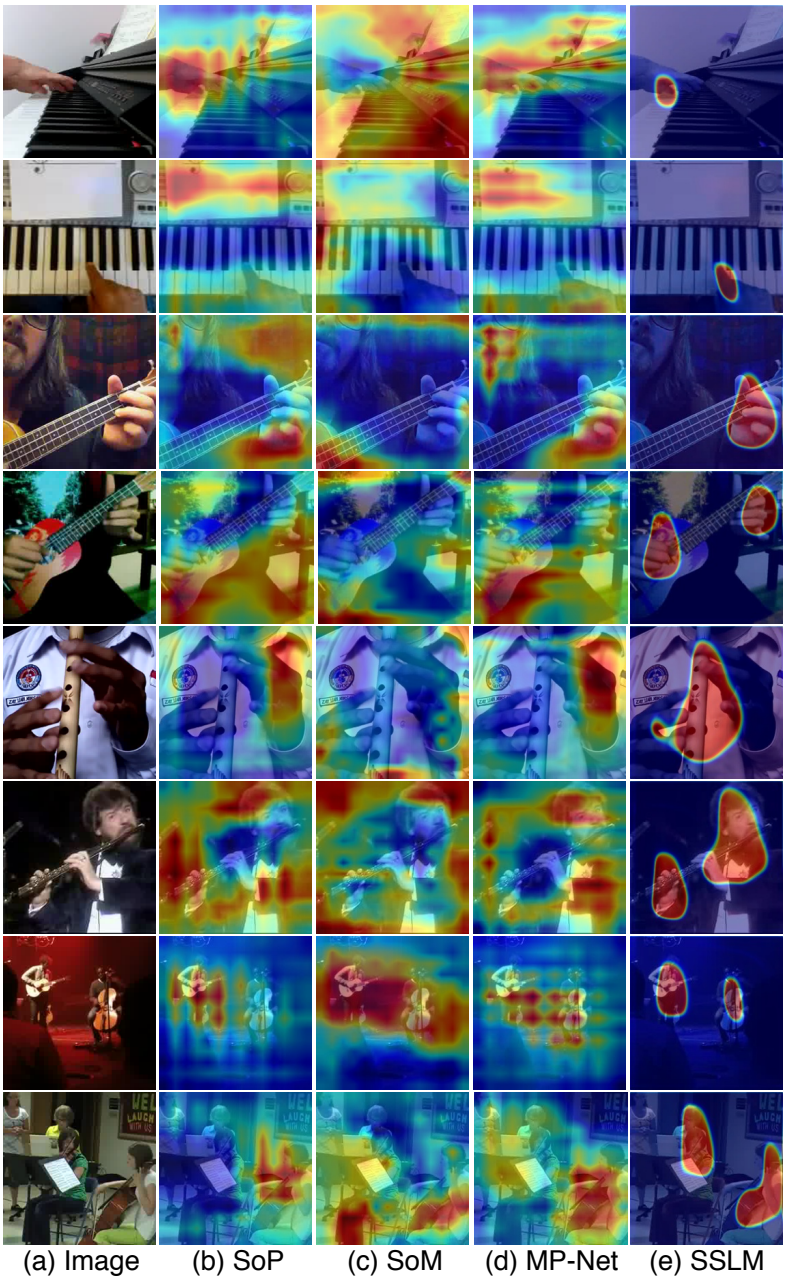

Fig. 5. Visualizing sound source location of our proposed SSLM network in comparison with baseline methods SoP, SoM, and MP-Net on A-MUSIC dataset.

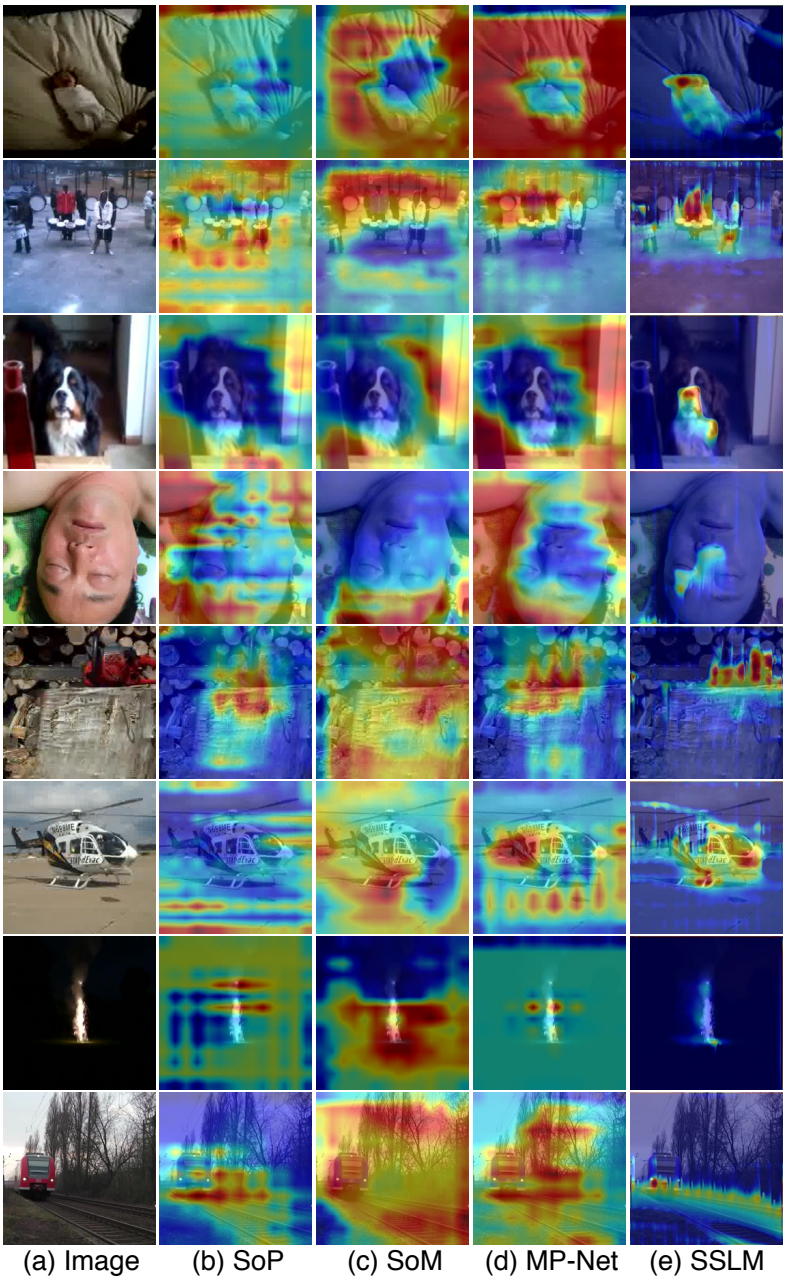

Fig. 6. Visualizing sound source location of our proposed SSLM network in comparison with baseline methods SoP, SoM, and MP-Net on A-NATURAL dataset.