Visually Guided Audio Source Separation with Meta Consistency Learning –Supplementary Materials–

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

S1. Qualitative Examples

We show additional qualitative results in addition to our main manuscript in Fig. S1. In the figure, the first and second rows show input mixed video pairs with associated mixed audio respectively. The 3^{rd} row shows the ground truth spectrograms and audio masks. The 4^{th}, 5^{th}, and 6^{th} rows present predicted spectrogram mask generated by different varients of our method. It is clear that the predicted mask and spectrogram generated by our final method (AVSS + CMC + Meta TTA) are the closest to ground truth out of all variants of the method..

S2. Additional Implementation Details

We set the decay parameter δ associated with λ in Eq. 2 of the main manuscript, using the following function from [2]: $\delta(\text{iter}) = \texttt{max}(0.1, 0.9^{\frac{\text{iter}}{100}})$ where *iter* refers to training iterations. We use batch size of 6 to train our meta-consistency models for 20 epochs, while we set the batch size to 1 during meta validation.

S3. Additional Experiments

S3.1. Cross-Modal Consistency Loss

We used euclidean distance (L2 norm) to compute the cross-modal consistency loss (Eq. 1) in the main manuscript. Here, we replace the euclidean distance with the cosine distance to validate our design choice. Table S1 summarizes the audio separation results under these variants. It is clear that we achieve higher performance under all the metrics using euclidean distance.

Method	SDR ↑	SIR↑	SAR↑
CMC (Cosine)	10.20	16.12	12.41
CMC (L2)	**10.35**	**16.71**	**12.43**

Table S1. The audio separation performance on MUSIC dataset with different loss types in cross-modal consistency loss.

LR, α	0	$1e^{-4}$	$1e^{-5}$	$2e^{-5}$	$1e^{-6}$	$1e^{-7}$
SDR	10.53	10.40	10.75	**10.80**	10.63	10.62
SIR	17.51	17.24	17.86	**17.87**	17.77	17.76
SAR	12.14	12.03	12.28	**12.30**	12.16	12.16

Table S2. Effects of the inner loop LR on MUSIC Val Set.

S3.2. Ablation Study

We conduct further ablation study to analyze various aspects of our proposed approach including the effects on inner loop learning rate (Sec. S3.2.1), the impact of different components (Sec. S3.2.2), and the sample-specific effects on the inner loop gradient updates (Sec. S3.2.3).

S3.2.1 Effects on inner loop learning rate

We further analyze the effects of various learning rates on the test-time adaptation performance. We believe that larger learning rates for the inner loop updates lead to degradation of model's performance even after a single update. In contrast, a lower learning rate can limit the capability of the model to adapt on unknown samples. To support this claim, we summarize the separation performance in terms of SDR, SIR, and SAR on MUSIC val set under different learning rates in Table S2. It is clear that we achieve the highest SDR, SIR, and SAR for the learning rate $2e^{-5}$, with a small gain in these metrics compared to the performance with $1e^{-5}$. However, the SDR performance is severely affected with higher learning rates (i.e., $1e^{-4}$).

S3.2.2 Network Architectures

In the main manuscript, we examined the importance of all components of our approach on MUSIC val set. Here, we conduct additional experiments on music test set to further analyze the components and summarize the results in Table S3. Our baseline AVSS achieves SDR of 10.62dB and SIR of 17.77dB without any audio-level class labels. When we include the cross-modal consistency loss

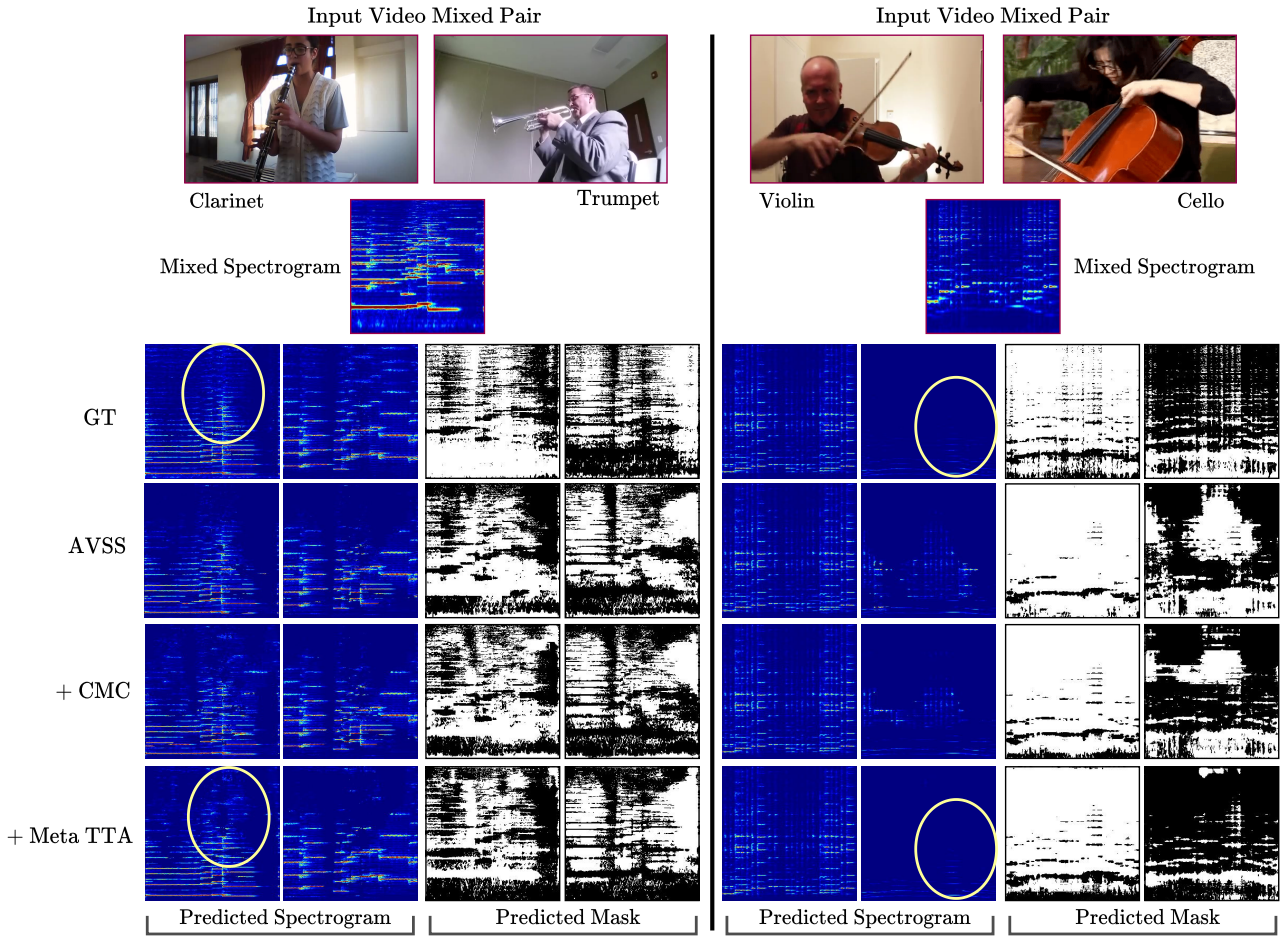

Figure S1. Qualitative audio separation results on the MUSIC21 test set for different variants of our proposed approach.

(CMC) with AVSS, AVSS+CMC marginally outperforms AVSS (10.62dB SDR vs 10.81dB SDR and 17.77dB SIR vs 17.80dB SIR). The meta-consistency training (AVSS + CMC + Meta) further promotes the separation performance substantially (11.38dB SDR and 18.74dB SIR). Finally, meta-consistency driven test-time adaptation (AVSS + CMC + Meta TTA) reasonably improves the overall separation performance (0.41dB SDR improvement). Interestingly, naive test-time adaptation (AVSS + CMC + Naive TTA) alone can not improve the separation performance which reveals the importance of our meta-consistency based training for source separation task. It is clear that the cross-modal consistency loss and meta-consistency training based test time adaptation promote the separation performance and we achieve the best results by employing both cross-modal consistency loss and meta-consistency training regarding all evaluation metrics.

Method	SDR ↑	SIR↑	SAR↑
AVSS	10.62	17.77	12.48
AVSS + CMC	10.81	17.80	12.53
AVSS + CMC + Meta	11.38	18.74	12.83
AVSS + CMC + Naive TTA	11.36	18.67	13.09
AVSS + CMC + Meta TTA	**11.77**	**19.36**	**13.13**

Table S3. The ablation results comparing different variants of our proposed pipeline on MUSIC test set. Our final method which incorporates both cross-modal consistency and meta-consistency training, outperforms all other baselines by a substantial margin.

S3.2.3 Effects on the sample-specific inner loop updates

In the main manuscript, we have shown that the meta-consistency learned models improve performance via test time adaptation for all the values of k used in meta-

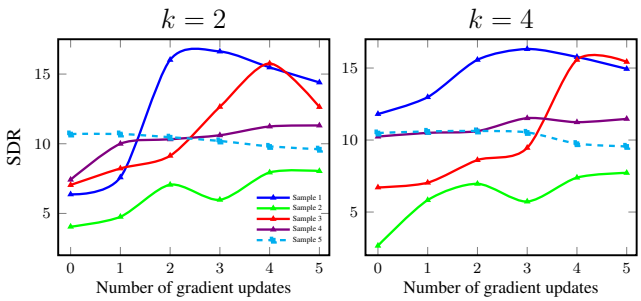

Figure S2. Illustration of SDR for randomly chosen test samples after each gradient updates for models trained with $k = \{2, 4\}$. Both plots show that test-time adaptation can improve performance in general, but the number of optimal gradient update steps may differ for each unseen music video. The dotted curves illustrate the samples for which test time adaptation even resulted in a slight performance degradation.

consistency training. We also observe that test time adaptation with smaller number of inner loop updates (i.e., $k = \{2, 3, 4\}$) shows the most SDR gain, while increasing the number of updates, k does not have any impact on improving separation performance. However, it was not clear if all the test samples significantly contribute towards the overall performance improvements. Recent work [1] pointed out that not all the test samples contribute equally to adaptation process even with meta-trained model, and in fact specific samples may lead to noisy gradients that could confuse the model after one gradient update. Figure S2 visually shows the adaptation process for five random test samples for which we show the performance for various k values ($k = 1, 2, 3, 4, 5$) using the meta-trained models with $k = \{2, 4\}$. The performance is improved with the increase of gradient updates for majority of samples, while the performance for few samples (e.g., dotted sample in Fig. S2) may degrade with a single gradient update. We further observe that the degree of improvement varies across different test samples and the number of gradient updates required to achieve peak SDR also varies.

S3.3. Differences Between MUSIC21 and MUSIC Dataset

In Table 3 of the main manuscript, we reported results under the setting where we first trained on MUSIC-21 dataset and evaluated on MUSIC dataset. MUSIC is a subset of MUSIC-21 dataset which has different number of classes and different data distribution. Therefore, a model trained on MUSIC-21 does not necessarily achieve optimal performance on the MUSIC dataset. To verify this, we evaluate the MUSIC-21 pretrained model directly on MUSIC dataset and report the results in Table S4. It is clear that fine-tuning on MUSIC dataset substantially improves the audio separation performance in terms of SDR, SIR, and SAR.

Method	SDR ↑	SIR↑	SAR↑
Ours (w/o fine-tune)	10.59	16.48	13.32
Ours (with fine-tune)	**12.81**	**19.56**	**14.16**

Table S4. The audio separation performance on MUSIC dataset before and after fine-tuning. Note that the results are for the AVSS+CMC variant of our approach.

References

[1] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022. 3

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