





# Multi-View Based Audio Visual Target Speaker Extraction

Peijun Yang <sup>1,2</sup>, Zhan Jin <sup>3</sup>, Juan Liu <sup>2,3,\*\*</sup>, Ming Li <sup>2,\*\*</sup>

<sup>1</sup>School of Cyber Science and Engineering, Wuhan University, Wuhan, China

<sup>2</sup>School of Artificial Intelligence, Wuhan University, Wuhan, China

<sup>3</sup>School of Computer Science, Wuhan University, Wuhan, China

liujuan@whu.edu.cn, ming.li.cuhksz@gmail.com

## Abstract

Audio-Visual Target Speaker Extraction (AVTSE) aims to separate a target speaker’s voice from a mixed audio signal using the corresponding visual cues. While most existing AVTSE methods rely exclusively on frontal-view videos, this limitation restricts their robustness in real-world scenarios where non-frontal views are prevalent. Such visual perspectives often contain complementary articulatory information that could enhance speech extraction. In this work, we propose Multi-View Tensor Fusion (MVTF), a novel framework that transforms multi-view learning into single-view performance gains. During the training stage, we leverage synchronized multi-perspective lip videos to learn cross-view correlations through MVTF, where pairwise outer products explicitly model multiplicative interactions between different views of input lip embeddings. At the inference stage, the system supports both single-view and multi-view inputs. Experimental results show that in the single-view inputs, our framework leverages multi-view knowledge to achieve significant performance gains, while in the multi-view mode, it further improves overall performance and enhances the robustness. Our demo, code and data are available at <sup>1</sup>

**Index Terms:** Audio-Visual Target Speaker Extraction, Robustness, Multi-View Tensor Fusion

## 1. Introduction

Target Speaker Extraction (TSE) separates a specific speaker’s voice from overlapping speech or background noise, making it essential for hearing aids and speech recognition systems. Traditional TSE approaches typically rely on auxiliary reference cues to direct attention toward the target speaker. Audio-only methods utilize the target speaker’s characteristics from a pre-recorded speech [1, 2, 3], while visually-guided approaches usually leverage the synchronized visual information [4, 5, 6] such as lip movements or auxiliary face, hence enrollment is not needed. Recently, visual cue based methods become more popular, as they do not rely on pre-registered information and remain unaffected by highly noisy acoustic interference. Most AVTSE systems follow an encoder-fusion-separator-decoder pipeline. Audio and visual encoders extract embeddings; the latter are upsampled and fused—typically via concatenation [7, 8, 9] or advanced strategies [10, 11]—before a separator network estimates a mask or target features, which is decoded into the target waveform. Common separators include GridNet [12], CrossNet [13], and DPRNN [14].

However, most existing visual-guided AVTSE systems predominantly assume the availability of frontal facial views. This

assumption is reflected in widely-used datasets such as LRS3 [15] and VoxCeleb2 [16], where speakers are typically captured from a frontal perspective. Since these datasets lack annotations for head pose variations, it remains unclear whether training and validation on stable frontal views is more effective than leveraging multiple distinct perspectives. Moreover, when synchronized multi-view video of a speaker is available, it raises the question of how to effectively exploit such synchronized information across different angles to further enhance model performance. Also, the reliance on frontal views significantly limits the practical applicability of AVTSE systems in real-world scenarios, where speakers naturally rotate their heads or cameras capture non-frontal angles—factors that have been shown to degrade performance [17].

Previous work generates pose-invariant information [18] or robust face frontalization [19, 20] for reducing the impact of head movements. Attempting to correct non-frontal views rather than learning to leverage it may discard the original visual information, which results in suboptimal performance when frontalization fails. Some works in other related fields attempt to identify and process each specific pose separately [21] or use multiple cameras’ information to enhance performance [22]. The former is limited by the number of poses, while the latter requires a fixed multi-camera setup in both training and testing. To address these limitations, we propose a novel multi-view learning framework. Instead of treating pose variation as a problem to be corrected via frontalization, we embrace it as a source of complementary information. The key insight is that synchronized multi-view videos inherently contain complementary articulatory cues—learning such cross-view correlations directly enhances each individual view’s representation, making it more robust to pose variations. Our framework learns a view-invariant and robust visual representation by explicitly modeling the multiplicative interactions between different viewing angles during training, rather than relying on the availability of multiple cameras at test time. Critically, this multi-view training strengthens the representation for each individual view. This is because the model learns to extract shared articulatory information across perspectives, enriching the feature space for each view rather than simply averaging or selecting among them. The model leverages multi-view data during training yet accepts single-view inputs at test time, making it practical for real-world single-camera scenarios. When tested on a video of a speaker freely turning their head, it applies learned cross-view knowledge to compensate for continuous pose changes—a capability frontalization methods cannot achieve.

\*\* indicates the corresponding author.

<sup>1</sup><https://b23ca07a.github.io/MVTF-Gridnet/>

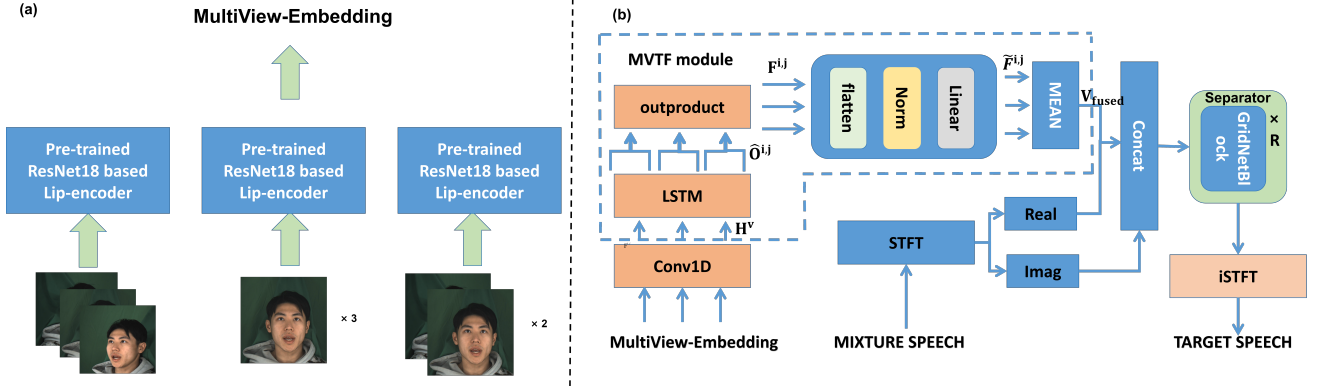


Figure 1: Our proposed MVTF method that extracts the target speakers’ speech using Multi-View lip embeddings. (a) Three views, one view (copied three times), and two views (one view copied two times) visible situations. (b) overall structure of MVTF-GridNet.

## 2. Methods

### 2.1. Overview

As shown in Fig1, we propose a robust AVTSE system that utilizes multi-view visual data in training to achieve superior performance during the inference even with a single (typically frontal) view testing video. Based on the TF-GridNet backbone [12], our model introduces a Multi-View Tensor Fusion module. The mixed audio is encoded by the Short-Time Fourier Transform (STFT), while lip encoders [23] process each video view. The MVTF module employs Long Short-Term Memory networks (LSTMs) [24] and a tensor outer product to fuse multi-view embeddings into a view-invariant representation. This is combined with audio features for the GridBlock to reconstruct the target speaker’s spectrogram, which is then converted to a waveform through iSTFT.

### 2.2. Audio Processing Backbone

We adopt TF-GridNet [12] as our audio separation backbone due to its strong performance in spectral mapping. The mixture audio signal  $\mathbf{X} \in \mathbb{R}^{B \times N \times M}$  (where  $B$  is the batch size,  $N$  is the number of waveform samples, and  $M$  is the number of microphones which defaults to 1) is first normalized by its standard deviation across channels and time to ensure stable training:  $\tilde{\mathbf{X}} = \mathbf{X} / \sigma(\mathbf{X})$ . A Short-Time Fourier Transform (STFT) encoder then projects the normalized waveform into a complex-valued spectrogram:  $\mathbf{C}_{mix} = \text{STFT}(\tilde{\mathbf{X}}) \in \mathbb{C}^{B \times M \times T_a \times F}$  where  $T_a$  and  $F$  represents the number of time frames and frequency bins, respectively. The real and imaginary components of  $\mathbf{C}_{mix}$  are concatenated to form the input feature map  $\mathbf{A} \in \mathbb{R}^{B \times 2 \times M \times T_a \times F}$  for the subsequent network.

### 2.3. Multi-View Feature Extraction

The multi-view video input  $\mathbf{V} \in \mathbb{R}^{B \times V \times T_v}$  (where  $B$  is the same batch size as the audio signal,  $V$  is the number of views and  $T_v$  is the number of video frames) is processed separately. For each view  $V_i \in \mathbb{R}^{B \times T_v}$ , we employ a pre-trained lip-reading network [23] to extract spatio-temporal features from the lip ROI sequence  $\mathbf{V}^v$ . Prior to feature extraction, an affine transformation is applied to align each ROI to a mean face template. The resulted embedding  $\mathbf{L}^v \in \mathbb{R}^{B \times T_v \times D}$  ( $D = 512$  here) provides a high-dimensional representation of the lip motion.

A critical challenge is the temporal misalignment be-

tween audio frames ( $T_a$ ) and video frames ( $T_v$ ). To address this, we apply linear interpolation to upsample the visual sequence to match the audio temporal resolution:  $\tilde{\mathbf{L}}^v = \text{Interpolate}(\mathbf{L}^v, \text{size} = T_a) \in \mathbb{R}^{B \times T_a \times D}$ . This ensures that each audio frame has a corresponding visual feature vector. Finally, we project the visual features into a common subspace shared with audio features using a 1D convolution layer:  $\mathbf{H}^v = \text{Conv1D}(\tilde{\mathbf{L}}^v) \in \mathbb{R}^{B \times T_a \times F}$ .

### 2.4. Multi-View Tensor Fusion Module

This module is designed to integrate information from multiple views even when some are missing during inference. In the event of missing perspectives during inference, our method replicates the available views to compensate for the absent ones. The features  $\mathbf{H}^v$  are fed into a shared, single-layer LSTM to capture temporal dependencies within each view’s sequence:

$$\mathbf{O}^v = \text{LSTM}(\mathbf{H}^v) \quad (1)$$

We use the output sequence  $\mathbf{O}^v \in \mathbb{R}^{B \times T_a \times F}$  for further fusion. We consider the lip embedding from different perspectives as multi-view but also shares a lot of similarity. Directly integrating multi-view inputs may introduce noise that degrades training. Inspired by the idea of the tensor fusion network [25] and the idea of fusing imperfect modality [26], we design the Multi-View Tensor Fusion Module to fuse the lip embedding of different views to suppress noise and enhance the performance of single view properly. To model rich multiplicative interactions between views—which additive concatenation is incapable of capturing—we compute pairwise outer products of the LSTM-processed features. First, we augment each feature vector with a constant 1 to incorporate bias terms. After constant 1 is appended, unimodal and bimodal interactions are all captured as described in [25]

$$\hat{\mathbf{O}} = [\mathbf{O}, \mathbf{1}] \in \mathbb{R}^{B \times T_a \times (F+1)} \quad (2)$$

The fused representation for a pair of views ( $i, j$ ) is computed as:

$$\mathbf{F}^{i,j} = \hat{\mathbf{O}}_i \otimes \hat{\mathbf{O}}_j \quad (3)$$

where  $\otimes$  denotes the outer product applied along the feature dimension for each aligned time step independently, resulting in  $\mathbf{F}^{i,j} \in \mathbb{R}^{B \times T_a \times (F+1) \times (F+1)}$ , which captures multiplicative interactions between the two view embeddings.

Table 1: Performance of MVTF-GridNet with single-view input and other fusion strategies on the MEAD dataset. “Training” denotes the lip embedding strategy used during training. For MVTF-GridNet, “Random” selects 3 distinct views out of 7 per batch; “Repeat” selects 1 view and repeats it three times; “Front” uses only frontal view repeated. For GridNet, “Random” selects 1 out of 7 views per batch. Simple Addition and Attention Fusion are trained with the “Random” strategy.

ID.	Model	Training	Metric	Testing View							AVG(7)
				Front	Top	Down	Left30	Left60	Right30	Right60	
-	Mixture	-	SI-SDR	-0.191							
1	MVTF-GridNet	Random (random 3 out of 7 views)	SI-SDR	<b>15.836</b>	<b>15.196</b>	<b>15.784</b>	<b>15.839</b>	<b>15.792</b>	<b>15.822</b>	<b>15.758</b>	<b>15.718</b>
2	MVTF-GridNet	Repeat (random 1 out of 7 views)*3	SI-SDR	15.133	14.878	15.202	15.108	15.138	15.146	15.109	15.102
3	MVTF-GridNet	Front *3	SI-SDR	14.930	10.230	14.841	14.778	14.456	14.761	14.723	14.102
4	GridNet	Random (random 1 out of 7 views)	SI-SDR	15.259	14.765	15.048	15.055	15.107	15.214	15.175	15.089
5	GridNet	Front	SI-SDR	13.290	7.731	11.159	13.622	13.321	13.810	13.914	12.406
-	MVTF-GridNet	Random (random 3 out of 7 views)	SDR <sup>†</sup>	10.784	10.646	10.865	10.873	10.971	10.801	10.740	10.811
-	PIAVE[18]	-	SDR <sup>†</sup>	11.773	8.923	8.514	8.583	6.118	8.387	4.935	8.176
6	Projected Addition	Random	SI-SDR	14.733	14.191	14.764	14.629	14.588	14.711	14.523	14.591
7	Attention Fusion	Random	SI-SDR	13.938	13.931	13.937	13.940	13.937	13.938	13.936	13.938

<sup>†</sup> Although we adopt the same dataset, data splits, mixture generation, and SNR range as PIAVE [18], utterance-level splits may differ. Furthermore, PIAVE is pre-trained on LRS3 and fine-tuned on MEAD, while our models are trained from scratch. Thus, results are indicative rather than strictly comparable.

The pairwise tensors  $\mathbf{F}^{i,j}$  are flattened and projected back to the original dimension  $F$  using a LayerNorm and a linear layer to maintain tractability.

$$\tilde{\mathbf{F}}_{i,j} = \text{Linear}(\text{LayerNorm}(\text{Flatten}(\mathbf{F}_{i,j}))) \quad (4)$$

The final fused visual context  $\mathbf{V}_{\text{fused}}$  is obtained by averaging the contributions from all available view pairs:

$$\mathbf{V}_{\text{fused}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \tilde{\mathbf{F}}^{i,j} \in \mathbb{R}^{B \times T_a \times F} \quad (5)$$

where  $\mathcal{P}$  is the set of all pairs of available views (The interaction between one view and itself is allowed so that we can use multi-view information to train and use only one view to infer). In our experiments,  $|\mathcal{P}|$  was set to 3 which means we have 3 available views and pairwise interactions produce 3 results, This tensor is unsqueezed to  $\mathbb{R}^{B \times 1 \times T_a \times F}$  to be combined with the audio features. Due to the rotation symmetry of the MVTF module, the order of inputs has no effect on the results. The advantage of the symmetry is that the camera positions in the test environment do not need to be the same as those in training and can be freely matched.

### 3. Experimental Setup

#### 3.1. Dataset

We conduct our experiments on the MEAD [27] dataset, a multi-view audio-visual emotional dataset. To isolate the impact of view variation from emotional variance, we exclusively use videos with neutral emotion for all speakers. The dataset provides synchronized audio and video from 7 different camera angles (front, top, down, left 30°, right 30°, left 60°, right 60°). For each mixture, the target speaker and an interfering speaker are randomly selected from different speakers in the dataset. All mixtures are a mix of two speakers.

A total of 10,000, 1,000, and 1,000 audio mixtures were generated for training, validation, and testing, respectively, ensuring that there are no overlap between the training, validation and testing partitions at either the utterance or speaker level. Their clean utterances are mixed at a random Signal-to-Noise Ratio (SNR) between -10 dB and 10 dB. All audio signals are resampled to 16 kHz. The video frames are processed at 25 frames-per-second(FPS).

In real-life scenarios, the head posture is constantly changing, so it is difficult to obtain stable single-view information. In order to verify the robustness of the model under real-life head posture changes, we construct a test set from the same 1,000 test mixtures by injecting feature segments from alternate viewpoints (e.g., 30° left) into frontal view sequences to test model’s robustness, while keeping the corresponding audio mixtures unchanged. The inserted segments varied in length from 20% to 40% of the total sequence and were placed randomly between the 30% and 80% positions of the timeline.

#### 3.2. Implementation Details

The proposed model was implemented using the PyTorch framework. The Adam [28] optimizer was employed for training with an initial learning rate of 1e-3. A reduction-based learning rate scheduler was utilized, which halved the learning rate if no performance improvement on the validation set was observed for three consecutive epochs. Training was terminated early if no improvement occurred for ten consecutive epochs. To ensure training stability, gradient clipping was applied, constraining the L2-norm of the gradients to a maximum value of 1. The models were trained for a maximum of 100 epochs and optimized using a Scale-Invariant Signal-to-Distortion ratio (SI-SDR) [29] loss.

#### 3.3. Baseline and Ablation Studies

To comprehensively evaluate the proposed MVTF module, we construct a hierarchical set of baselines and ablations. All models share the same TF-GridNet backbone, optimization settings, and evaluation protocol. The comparisons span two dimensions: the presence of the MVTF module and the input strategy during training, as detailed below.

We first establish two MVTF-free baselines: GridNet (Front), trained solely on frontal views, and GridNet (Random), trained on random single views per batch to isolate the effect of viewpoint diversity. Next, to assess the contribution of the MVTF module itself, we train MVTF-GridNet under three input strategies: Random (three distinct and random views out of seven views per batch), Repeat (one random view out of seven views per batch and repeat it three times), and Front 3 (only frontal view repeated). Finally, to validate the effectiveness and necessity of the proposed MVTF module, we compare MVTF against two alternative fusion strategies under the

same multi-view training setup: Projected Addition (element-wise summation after linear projection, capturing only additive interactions) and Attention Fusion (crossattention with random QKV assignment, offering flexibility but potential instability).

All models are evaluated on the MEAD test set with single-view inputs. For MVTF-GridNet, the singleview inputs are repeated 3 times to meet the number of input channels. Results are summarized in Table 1

## 4. Results and Discussion

The performance of all systems was quantitatively evaluated using Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [29], Perceptual Evaluation of Speech Quality (PESQ) [30] and the Short-Term Objective Intelligibility (STOI) [31].

### 4.1. Multi-View Training Enhancement for Single-View Testing

As shown in Table 1, MVTF-GridNet trained on random multi-view data achieves the best average SI-SDR of 15.718 dB, outperforming its frontal-only counterpart by 1.616 dB. The gain is especially significant under challenging views such as the top view, demonstrating the method’s effectiveness in handling non-frontal perspectives. The performance gap between Random (ID 1) and Repeat (ID 2) or Front (ID 3) strategies imply that exposure to diverse viewpoints during training is critical for generalization. This is also confirmed by the fact that all models trained with random viewpoints perform better than those trained with only frontal faces. Interestingly, GridNet trained only on frontal views (ID 5) performs worse on frontal than on some profile views (Right30/60), may be attributed to the pre-trained visual encoder’s inherent robustness to certain angles. However, MVTF-GridNet exhibits superior robustness.

Table 2: *Robustness to mixed frontal/non-frontal views. Models (ID 1, 4, 5) correspond to MVTF-GridNet (Random), GridNet (Random), and GridNet (Front) from Table 1, respectively.*

Model	SI-SDR	PESQ	STOI
MVTF-GridNet (ID 1)	15.834	2.887	0.915
GridNet-Random (ID 4)	15.179	2.778	0.902
GridNet-Front (ID 5)	10.425	2.470	0.851

When evaluated on mixed-view sequences simulating head rotations (Table 2), MVTF-GridNet (ID 1) maintains stable performance, while GridNet (ID 4,5) shows degradation—especially ID 5. This indicates that MVTF effectively learns cross-view correlations, improving stability over naive multi-view or single-view training.

### 4.2. Performance for Multi-View Conditions during Inference

Table 3 demonstrates the inference performance of typical multi-view combinations in real-world settings, underscoring the robustness and adaptability of our method. All configurations consistently achieve high performance, with SI-SDR around 15.85 dB, PESQ near 2.90, and STOI consistently above 0.915. This consistency indicates that the model performs reliably across various angle arrangements without dependence on a specific configuration. Such flexibility is particularly beneficial in practical scenarios where multiple cameras are fixed while the target speaker moves.

Table 3: *Performance of typical multi-view input combinations. The test model is ID 1 of table 1. The test set is the same 1,000 test mixtures by using the corresponding view combination .*

Combination	SI-SDR	PESQ	STOI
Front&Left30&Right30	15.857	2.898	0.915
Front&Left60&Right60	15.854	2.898	0.916
Front&Right30&Right30	15.834	2.892	0.915
Front&Right60&Right60	15.835	2.894	0.915
Front&Left30&Right60	15.851	2.899	0.916

### 4.3. Comparison of Fusion Strategies and Model Complexity

Table 4: *Model parameter count and computational cost (4-second mixture input).*

Model	Params (M)	FLOPs (G)	SI-SDR (dB)
GridNet (Single-View)	7.235	470.750	15.089
MVTF-GridNet (Ours)	7.561	471.824	15.718
Projected Addition	7.274	470.934	14.591
Attention Fusion	7.269	470.921	13.938

We compare MVTF with Projected Addition and Attention Fusion under identical settings (ID 6–7). MVTF achieves the highest SI-SDR among the three. Both alternatives underperform relative to the random single-view GridNet baseline: Projected Addition fails to capture nonlinear interactions, while Attention Fusion suffers from unstable cross-attention due to random QKV assignment. This highlights the challenge of fusing multi-view information without introducing noise. In contrast, MVTF explicitly models multiplicative interactions via outer products, enabling more effective fusion. MVTF-GridNet adds only marginal parameters and FLOPs over the single-view baseline (Table 4), yet yields substantial gains—striking an excellent trade-off between complexity and performance.

### 4.4. Comparison with previous works on the MEAD dataset

We further compare with PIAVE [18], a recent AVTSE model that generates pose-invariant frontal faces. As shown in Table 1, our MVTF-GridNet achieves an average SDR of 10.81 dB across seven views, outperforming PIAVE’s 8.18 dB. This gap underscores the effectiveness and stability of our explicit multi-view fusion mechanism combined with the TF-GridNet[12] backbone for target speaker extraction under varying viewpoints.

## 5. Conclusion

This paper presents MVTF, a novel Multi-View Tensor Fusion framework for AVTSE. By explicitly modeling multiplicative interactions across different viewpoints, our method captures complementary articulatory information from multi-view lip videos during training, enabling robust performance with both multi-view and single-view inputs at test time. Experimental results on the MEAD dataset demonstrate that MVTF outperforms frontal-only and naive multi-view baselines, particularly on challenging non-frontal views. The framework’s flexible input configuration makes it well-suited for real-world scenarios with varying camera angles.

## 6. Acknowledgements

This work was supported by the National Key Research and Development Program of China (1502-211171515).

## 7. Generative AI Use Disclosure

Generative AI tools were used for limited language editing purpose, including improving clarity and correcting grammatical issues. No substantive content, analysis, or conclusions were generated by AI. The authors remain fully responsible for the content of this manuscript.

## 8. References

- [1] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.
- [2] F. Hao, X. Li, and C. Zheng, "X-tf-gridnet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion," *Information Fusion*, vol. 112, p. 102550, 2024.
- [3] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Proc. Interspeech 2020*, 2020, pp. 1406–1410.
- [4] Z. Pan, R. Tao, C. Xu, and H. Li, "Muse: Multi-modal target speaker extraction with visual cues," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6678–6682.
- [5] T. Alfouas, J. Chung, and A. Zisserman, "The conversation: deep audio-visual speech enhancement," in *Interspeech*, vol. 2018. International Speech Communication Association, 2018.
- [6] Z. Pan, G. Wichern, Y. Masuyama, F. G. Germain, S. Khurana, C. Hori, and J. Le Roux, "Scenario-aware audio-visual tf-gridnet for target speech extraction," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [7] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 667–673.
- [8] V. A. Kalkhorani, C. Yu, A. Kumar, K. Tan, B. Xu, and D. Wang, "Av-crossnet: an audiovisual complex spectral mapping network for speech separation by leveraging narrow-and cross-band modeling," *IEEE Journal of Selected Topics in Signal Processing*, 2025.
- [9] Z. Pan, M. Ge, and H. Li, "Usev: Universal speaker extraction with visual cue," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3032–3045, 2022.
- [10] K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu, "An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6637–6651, 2024.
- [11] K. Li, R. Yang, F. Sun, and X. Hu, "Iianet: An intra-and inter-modality attention network for audio-visual speech separation," *arXiv preprint arXiv:2308.08143*, 2023.
- [12] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation," in *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [13] V. A. Kalkhorani and D. Wang, "Tf-crossnet: Leveraging global, cross-band, narrow-band, and positional encoding for single-and multi-channel speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4999–5009, 2024.
- [14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.
- [15] T. Afouas, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
- [16] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.
- [17] T. R. Jordan and K. Bevan, "Seeing and hearing rotated faces: influences of facial orientation on visual and audiovisual speech recognition," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 2, p. 388, 1997.
- [18] Q. Liu, M. Ge, Z. Wu, and H. Li, "Piave: A pose-invariant audio-visual speaker extraction network," in *Proc. Interspeech 2023*, 2023, pp. 3719–3723.
- [19] Z. Kang, M. Sadeghi, R. Horaud, X. Alameda-Pineda, J. Donley, and A. Kumar, "The impact of removing head movements on audio-visual speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7302–7306.
- [20] Z. Kang, R. Horaud, and M. Sadeghi, "Robust face frontalization for visual speech recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2485–2495.
- [21] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel, "Multi-view facial expression recognition using local appearance features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 3533–3536.
- [22] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *British Machine Vision Conference, BMVC 2017*, 2017.
- [23] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.
- [24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [25] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.
- [26] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1569–1576.
- [27] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European conference on computer vision*. Springer, 2020, pp. 700–717.
- [28] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *ICLR: international conference on learning representations*, 2015, pp. 1–15.
- [29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.
- [30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.