CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

AVCLNet: Multimodal Multispeaker Tracking Network Using Audio-Visual Contrastive Learning

Yihan Li[1] | Yidi Li[1,2] | Zhenhuan Xu[1] | Hao Guo[1] | Mengyuan Liu[3,4] | Weiwei Wan[2]

[1]College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China | [2]Graduate School of Engineering Science, The University of Osaka, Osaka, Japan | [3]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen, China | [4]Peking University, Shenzhen Graduate School, Shenzhen, China

Correspondence: Yidi Li (liyidi@tyut.edu.cn) | Zhenhuan Xu (xuzhenhuan@tyut.edu.cn)

Received: 10 April 2025 | **Revised:** 5 August 2025 | **Accepted:** 16 October 2025

Keywords: computer vision | machine perception | multimodal approaches | pattern recognition | video signal processing

ABSTRACT

Audio-visual speaker tracking aims to determine the locations of multiple speakers in the scene by leveraging signals captured from multisensor platforms. Multimodal fusion methods can improve both the accuracy and robustness of speaker tracking. However, in complex multispeaker tracking scenarios, critical challenges such as cross-modal feature discrepancy, weak sound source localisation ambiguity and frequent identity switch errors remain unresolved, which severely hinder the modelling of speaker identity consistency and consequently lead to degraded tracking accuracy and unstable tracking trajectories. To this end, this paper proposes a multimodal multispeaker tracking network using audio-visual contrastive learning (AVCLNet). By integrating heterogeneous modal representations into a unified space through audio-visual contrastive learning, which facilitates cross-modal feature alignment, mitigates cross-modal feature bias and enhances identity-consistent representations. In the audio-visual measurement stage, we design a vision-guided weak sound source weighted enhancement method, which leverages visual cues to establish cross-modal mappings and employs a spatiotemporal dynamic weighted mechanism to improve the detectability of weak sound sources. Furthermore, in the data association phase, a dual geometric constraint strategy is introduced by combining the 2D and 3D spatial geometric information, reducing frequent identity switch errors. Experiments on the AV16.3 and CAV3D datasets show that AVCLNet outperforms state-of-the-art methods, demonstrating superior robustness in multispeaker scenarios.

1 | Introduction

Multispeaker tracking is a critical task in the field of human–robot interaction, aiming to determine the spatial locations and identity associations of multiple speakers in complex scenarios by analysing real-time data from multimodal sensors such as microphone arrays and cameras [1]. This technology has significant applications in multiparty video conferencing systems, intelligent group behaviour analysis and social navigation for service robots [2]. Tracking problems are typically addressed using computer vision-based object tracking methods [3–5] and auditory-based sound source localisation (SSL) methods [6, 7]. However, unimodal approaches face significant challenges in multitarget tracking. Visual-based methods are susceptible to dense occlusions, sudden illumination changes and cross-view identity switches, whereas acoustic-based methods suffer from speech overlapping when multiple speakers talk simultaneously and are highly sensitive to noise and reverberant environments. To address these limitations, tracking frameworks that integrate multimodal perception have become a key pathway to enhancing tracking robustness. By associating acoustic features from audio streams with speaker facial representations from visual streams, a cross-modal temporal association model can be constructed, significantly

improving tracking accuracy and continuity in multiparty interaction scenarios.

Currently, existing audio-visual multimodal tracking networks primarily focus on feature fusion and cross-modal association during the fusion phase. For instance, prior studies employ early or late fusion strategies to integrate audio-visual information, typically via direct feature concatenation or weighted combination [8–10], but these methods have limited adaptability to heterogeneous modalities. With the advent of attention mechanisms, the multimodal perception attention network employs a self-supervised cross-modal strategy to assess the significance of visual and auditory measurements, enabling weighted fusion of the multimodal features [11]. The cross-modal multihead attention mechanism is introduced to facilitate the interaction of information from the visual and audio streams [12, 13]. However, in multiperson interactions, such as when multiple speakers are spatially close or speaking simultaneously, the association between visual and auditory features may be incorrect, leading to an identity switch. It is worth noting that recent studies have attempted to enhance multimodal representation capabilities through self-supervised audio-visual contrastive learning mechanisms. For example, AV-HuBERT [14] utilises contrastive and masked prediction strategies to construct token-level alignment between speech and lip movements, achieving good performance in tasks such as speech recognition and lip-reading. However, AV-HuBERT [14] is primarily designed for semantic understanding tasks, with its feature optimisation objectives focused on content-level matching. It lacks modelling of speaker identity and spatial awareness, making it difficult to directly adapt to multispeaker tracking scenarios, which require identity consistency and trajectory continuity. To address this challenge, this paper proposes an audio-visual contrastive learning (AVCL) mechanism, which integrates audio-visual features into a unified feature space to overcome feature bias caused by modality heterogeneity in traditional methods. Unlike contrastive mechanisms that focus solely on content-level speech consistency, AVCL introduces a task-driven positive and negative sample construction strategy. By pulling closer the positive sample pairs (audio-visual features of the same target), the network learns multimodal consistent representations of the same target, enabling finer-grained cross-modal feature alignment. Simultaneously, by pushing apart the negative sample pairs (audio-visual features of different targets), the network enhances its ability to discriminate between different targets, thereby improving identity-consistent representation and reducing identity switch errors in dense multispeaker interaction scenarios. In addition, the AVCL module is directly integrated into the backbone of the multispeaker tracking network and jointly optimised with the feature extraction and data association modules, enabling unified learning of semantic modelling, spatial localisation and identity consistency, which significantly enhances the robustness and accuracy of multitarget tracking.

Feature extraction is a critical phase in audio-visual multimodal tracking networks, typically involving both acoustic and visual measurements. Among these, the generalised cross-correlation with phase transform (GCC-PHAT) [15] and the multiple signal classification (MUSIC) algorithm [16] are two of the most widely used acoustic measurement methods. A spatial-temporal global coherence field (stGCF)-based acoustic measurement method was proposed [11], which constructs spatial sampling points using a camera model and derives the optimal values over a period of time based on motion continuity. However, the peak distribution of stGCF is interfered with by multiple synchronous sound sources, resulting in poor performance in multisource scenarios. With the advancement of deep neural networks, data-driven direction-of-arrival (DOA) estimation methods based on deep learning have gained significant attention, such as contrastive learning-based multitarget DOA estimation under low signal-to-noise ratio (SNR) conditions [17] and lightweight deep neural network approaches that incorporate data redundancy removal and regression techniques [18], which better address the challenges in complex multisource scenarios. However, these methods lack the ability to locate weak sound sources in multisource environments and fail to effectively utilise spatial and temporal information. Therefore, this paper designs a vision-guided weak sound source weighted enhancement method, which uses visual cues to establish cross-modal mapping and weighted summation of historical acoustic cues to improve the detectability of weak audio signals.

Data association represents a critical challenge in audio-visual multiobject tracking networks, particularly during the trajectory tracking phase. This task becomes especially demanding in complex scenarios where precise cross-frame target matching is essential for maintaining tracking continuity. The nearest neighbour (NN) algorithm [19] is the simplest data association method, associating the closest detection values with target trajectories. However, the NN algorithm is prone to mistakenly associating irrelevant targets with the trajectory, leading to tracking errors. Joint probabilistic data association (JPDA) [20] is a probabilistic-based multitarget data association method; it selects the optimal association by computing the joint probability of all possible data associations. However, it requires prior knowledge of the number of targets and fails in the absence of targets. Multiple hypothesis tracking (MHT) [21] is an algorithm that generates multiple hypotheses during tracking, improving tracking robustness by evaluating different hypothesis paths. However, its high computational cost makes it difficult to adapt to complex target motion patterns, resulting in tracking drift. To address this challenge, this paper proposes a dual geometric constraint strategy that integrates both 2D and 3D spatial geometric information, effectively optimising the data association process through complementary constraint mechanisms. Traditional 2D geometric constraints consider only information in the image plane and cannot capture the actual position and motion of targets in 3D space. For instance, two targets may be very close in the image, but their actual distance in 3D space could be far apart. 3D geometric constraints rely on the position of targets in 3D space but ignore the target's appearance and posture in the image. The proposed multiple constraints compensate for the information loss caused by single constraints, enabling more precise capture of target appearance features and spatial positions, thereby effectively reducing identity switch rate and improving the accuracy of data association.

The contributions of this paper are summarised as follows:

- A novel multimodal multispeaker tracking network using audio-visual contrastive learning (AVCLNet) is proposed.

By integrating audio-visual features into a unified space and employing contrastive learning, the network enhances multimodal alignment and identity consistency, significantly improving tracking robustness in multispeaker interaction scenarios.

- A vision-guided weak sound source enhancement method is designed for multisource localisation, which establishes cross-modal audio-visual mapping through visual cue guidance. By implementing a spatiotemporal weighting strategy to integrate historical acoustic information, this approach significantly improves the detection capability of weak sound sources.
- A dual geometric constraint strategy is proposed for data association, integrating both 2D and 3D spatial information to optimise the cost matrix. This approach enhances the precision of target appearance and spatial feature matching, significantly improving association robustness.
- The proposed AVCLNet achieves superior performance over state-of-the-art methods on AV16.3 and CAV3D benchmarks, with significant improvements across key multiobject tracking metrics.

2 | Related Work

2.1 | Audio-Visual Speaker Tracking

The complementary nature of visual and auditory modalities effectively overcomes the inherent limitations of unimodal perception. This makes multimodal information integration a crucial approach for improving tracking performance. In audio-visual speaker tracking systems, visual measurements primarily rely on either handcrafted appearance features [22–28] or deep neural network-based discriminative features [29, 30]. On the auditory side, sound source localisation (SSL) algorithms, such as time delay estimation and direction-of-arrival (DOA) estimation [9, 31–35], have been widely applied in azimuth estimation. Early audio-visual speaker tracking methods were predominantly based on Bayesian frameworks. Particle filter (PF) [36–41] enables target distribution inference through nonlinear modelling but suffers from high computational complexity. Probability hypothesis density (PHD) filters [22, 42] allow for dynamic variations in the number of targets but struggle to distinguish similar targets in multispeaker tracking. The Poisson multi-Bernoulli mixture (PMBM) filter [32, 43] integrates a phase-aware strategy to enhance tracking in intermittent speech scenarios but relies on manually designed observation likelihood functions, limiting its adaptability to complex real-world environments. In recent years, deep learning methods have been widely applied in multimodal fusion research [44–46]. A cross-modal attention fusion mechanism was proposed to capture temporal dependencies within each modality and achieve alignment across modalities [13]. A cross-modal multihead cross-attention mechanism was proposed to jointly model multimodal context and interactions while incorporating a quality-aware module for multispeaker tracking [12]. However, these methods often lack identity management mechanisms, making them prone to identity switches. To address these challenges, the proposed AVCLNet constructs a unified multimodal feature space through an audio-visual contrastive learning mechanism to optimise cross-modal feature alignment and enhance identity-consistent representation. Additionally, by integrating a dual geometric constraint strategy, AVCLNet achieves joint optimisation of feature representation and the tracking model, demonstrating superior tracking performance in multispeaker interaction scenarios.

2.2 | Multisource Localisation

Research on multisource localisation has made significant progress, driven by advancements in both traditional signal processing methods and deep learning techniques. Traditional methods, such as generalised cross-correlation with phase transform (GCC-PHAT) [15] and steered response power with phase transform (SRP-PHAT) [47], demonstrate excellent performance in single-source scenarios but face limitations in multisource environments and reverberant conditions. To address these challenges, researchers have proposed the probabilistic graph diffusion model for the source localisation method [48], which introduces a probabilistic graph diffusion model to enhance the accuracy of source localisation. In recent years, deep learning-based approaches have led to breakthroughs in multisource localisation. Convolutional recurrent neural networks (CRNN) [49] integrate time-frequency feature extraction with sequential modelling capabilities, achieving high-precision localisation in noisy environments. With the development of transformer architectures, a self-attention mechanism was proposed to enhance long-term dependency modelling [50] while further incorporating a multihead self-attention mechanism to improve localisation robustness in complex acoustic environments [51]. Additionally, traditional spatial features are combined with deep learning models [52, 53], explicitly leveraging the physical properties of sound fields to mitigate the impact of reverberation on localisation accuracy. However, existing multisource localisation methods primarily rely on audio information, lacking collaborative optimisation with the visual modality. This leads to increased source confusion in scenarios involving overlapping speakers. Moreover, current methods do not incorporate effective enhancement mechanisms for weak sound sources, making them susceptible to masking in noisy and reverberant environments. In this paper, we propose a vision-guided weak sound source weighted enhancement method. By establishing a cross-modal mapping, the method leverages visual cues to guide auditory measurements and employs weighted aggregation of historical acoustic cues to enhance the detectability of weak sound sources, thereby improving the robustness of multisource localisation.

2.3 | Multimodal Contrastive Learning

Research on multimodal contrastive learning has primarily focused on cross-modal alignment and representation optimisation between vision and language modalities. For instance, CLIP [54] leverages large-scale image-text contrastive learning to achieve open-domain generalisation, whereas ALIGN [55] demonstrates the robustness of cross-modal contrastive learning in noisy data scenarios. Additionally, VLMo [56] further

enhances modality interaction diversity and task adaptability through architectural innovations and decoupling strategies. Regarding audio-visual contrastive learning, most existing studies primarily adopt self-supervised paradigms, leveraging contrastive learning to align cross-modal representations. For example, HiCMAE [57] introduces a hierarchical contrastive masked autoencoder, applying contrastive learning between masked and unmasked views across modalities to capture high-level semantic correlations, leading to strong performance in emotion recognition. SCAV [58] models the sequential structure of audio-visual signals by performing contrastive learning over nonaggregated representations, thereby enforcing temporal consistency via sequence-wise distance. Furthermore, DETECLAP [59] enhances audio-visual representation learning by introducing an object-aware audio-visual tag prediction loss, improving performance on retrieval and classification benchmarks. AV-HuBERT [14], as a powerful self-supervised framework, learns audio-visual speech representations via masked modelling and clustering mechanisms. In contrast, the proposed audio-visual contrastive learning strategy goes beyond simple cross-modal alignment by constructing identity-aware positive and negative sample pairs. It explicitly optimises the feature fusion of the same speaker across heterogeneous modalities while simultaneously enhancing the discriminability between different speakers, thereby significantly improving identity consistency modelling in speaker tracking tasks.

3 | Methodology and Network Design

In this paper, we propose a novel audio-visual contrastive learning-based network, AVCLNet, for multimodal multi-speaker tracking. As illustrated in Figure 1, the framework consists of three stages: audio-visual measurement, audio-visual contrastive learning and data association. First, the input audio-visual signals are composed of temporally synchronised audio-visual sample pairs (V_t, A_t), where $t = 1, ..., T$. T represents the total number of frames. In the audio-visual measurement stage, a face detector is employed to extract visual cues O_t, whereas a vision-guided audio measurement is adopted to obtain enhanced sound source maps M_t. The audio and visual cues are encoded by the audio and visual encoder, generating audio and visual feature representations I_t and S_t. In the audio-visual contrastive learning stage, the extracted audio and visual features are aligned and optimised through a contrastive learning mechanism. Subsequently, an attention mechanism fuses the modalities to generate audio-visual representations F_t, which are further processed by a prediction head to produce detection results $\hat{p}_t$. Finally, in the data association stage, a dual geometric constraint strategy is introduced to associate cross-frame detection results, obtaining speaker motion trajectories Tr_t. In this section, we provide a detailed description of the AVCLNet tracking network.

3.1 | Audio-Visual Measurement

3.1.1 | Visual Measurement

In this stage, a lightweight face detection model $f_v^{\text{det}}(\cdot)$ is employed for real-time face detection, which processes the input video stream in an end-to-end manner using convolutional neural networks. The detection process can be formalised as follows:

$$O_t = f_v^{\text{det}}(V_t), \tag{1}$$

where $V_t \in \mathbb{R}^{H\times W\times 3}$ represents the current input image frame, $O_t = \{o_t^i\}_{i=1}^{N_v}$ is the set of all detected face bounding boxes at frame t, $i \in \{1, 2, ..., N_v\}$ represents the speaker ID, and N_v is the number of detected faces. $o_t^i = (u_t^i, v_t^i, w_t^i, h_t^i)^\top$ denotes the face

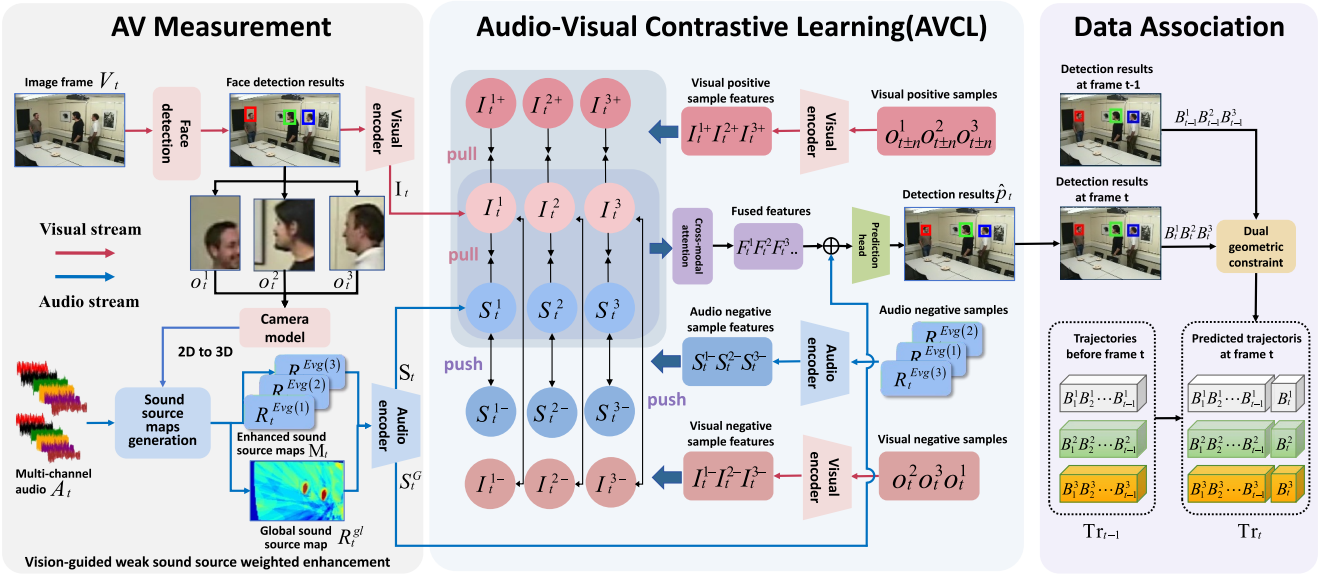

FIGURE 1 | The overall framework of AVCLNet comprises three main stages: audio-visual measurement, audio-visual contrastive learning and data association. First, facial regions are obtained through visual detection, and a visually guided audio measurement is employed. Then, cross-modal feature alignment is optimised through audio-visual contrastive learning (AVCL). Finally, a dual geometric constraint strategy is applied for data association.

bounding box coordinates of the speaker with ID = i, where (u_t^i, v_t^i) represents the top-left corner coordinates of the bounding box, and (w_t^i, h_t^i) corresponds to its width and height. Visual cues O_t are then fed into the visual encoder to extract deep visual features, which can be formulated as follows:

$$I_t = f_v^{enc}(O_t), \tag{2}$$

where $I_t = \{I_t^i\}_{i=1}^{N_t}$ is the set of encoded visual features, and I_t^i denotes the visual feature of the speaker i at frame t.

3.1.2 | Audio Measurement

In this stage, the multichannel audio A_t is processed using stGCF-based audio measurement [11] to generate the sound source map, where the peak distribution indicates the location of the sound source. However, in multisource scenarios, the distribution is interfered with environmental noise and reverberation. Therefore, when reliable visual observations are available, we use visual cues to guide audio measurement by narrowing the spatial sampling range based on the results of face detection, thus avoiding a global search of the entire image frame. This vision-guided approach effectively reduces noise interference from sources in other regions. Additionally, when speakers have weak or no speech at certain moments, the peaks in the stGCF-based sound source map often fail to indicate the true target location. To address the issue of weak sound sources being affected by noise interference, we propose a vision-guided weak sound source weighted enhancement method, which enhances weak sound sources by a weighted combination of historical sound source maps. This method improves the detection strength of weak sound sources in both the temporal and spatial dimensions, effectively improving weak sound source localisation performance in multisource localisation scenarios. The detailed process is shown in Figure 2.

In order to generate a global sound source map based on stGCF, a set of 2D sampling points $\{(x, y)\}$ is first extracted from the entire image frame to represent candidate pixels of potential sound source locations. Each 2D sampling point is projected onto multiple predefined depths d using a calibrated pinhole camera model, resulting in 3D projection sampling points $(x, y, d) = \Phi((x, y), d)$, where $\Phi(\cdot)$ is the back-projection function defined by the intrinsic and extrinsic parameters of the camera. A structured 3D sampling grid is finally formed across spatial and depth dimensions, as illustrated in Figure 3. The generalised cross-correlation with phase transform (GCC-PHAT) is used to measure the coherence between audio signals received by multiple microphones. Based on GCC-PHAT, the global coherence field (GCF) value is calculated at the sampling grid locations, generating a global coherence field (GCF) map. Then, the depth at which the peak of the GCF map is selected is followed by the selection of the GCF map with the maximum peak over a segment of time, generating the global sound source map, denoted as R_t^{gl}.

If a face is detected in the visual measurement, the face region is extracted to constrain the sound source localisation, generating a vision-guided sound source map R_t^{vg}. First, a 3D sampling grid is generated within the detected face bounding box O_t. Then, the GCF value G_t^i is calculated at each sampling point in the 3D space. The spatial parameters of the weak sound source enhancement weight are defined as follows:

$$\Theta_t^{\max(i)} = \max\{G_t^i(x, y, d)\}, \tag{3}$$

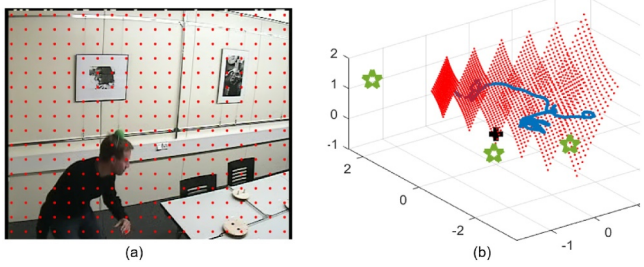

FIGURE 3 | The 2D sampling points (a) and the corresponding 3D projection sampling points (b).

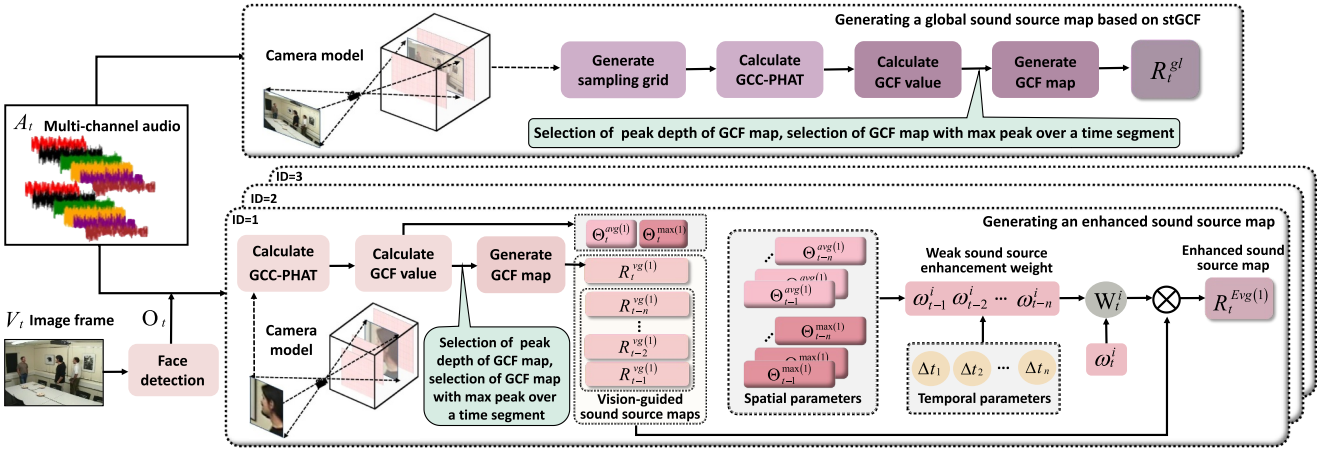

FIGURE 2 | The schematic diagram of the vision-guided weak sound source weighted enhancement module. This module utilises a camera model combined with face detection results to generate vision-guided sound source maps and performs weighted aggregation of historical information through spatiotemporal parameters to generate enhanced sound source maps.

$$\Theta_t^{avg(i)} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} G_t^i(x, y, d), \tag{4}$$

where (x, y, d) represents the sampling point in the 3D space ($x \in \{1, 2, ..., w\}$, , and the spatial parameters $\Theta_t^{max(i)}$ and $\Theta_t^{avg(i)}$ respectively represent the maximum GCF value and average GCF value for the speaker i at frame t. The weak sound source enhancement weight W_t^i is jointly determined by the spatial parameters and temporal parameters Δt_m, and is calculated as follows:

$$\omega_{t-m}^i = \alpha \cdot \Theta_{t-m}^{\max(i)} \cdot \Theta_{t-m}^{avg(i)} \cdot \frac{1}{\Delta t_m}, m = 1, 2, ..., n, \tag{5}$$

$$W_t^i = \left[\omega_t^i, \omega_{t-1}^i, \omega_{t-2}^i, ..., \omega_{t-n}^i\right], \omega_t^i = 0.5, \sum_{m=1}^{n} \omega_{t-m}^i = 0.5, \tag{6}$$

where α represents the normalisation factor, and $\Delta t_m = m$ represents the time difference between the historical frame and the current frame. For the vision-guided sound source map R_t^{vg} generated from the face region, the enhanced sound source map is obtained by weighting with weight W_t^i:

$$R_t^{Evg(i)} = \left[R_t^{vg(i)}, R_{t-1}^{vg(i)}, R_{t-2}^{vg(i)}, ..., R_{t-n}^{vg(i)}\right]^\top \otimes W_t^i, \tag{7}$$

where $R_{t-n}^{vg(i)}$ represents the vision-guided sound source map generated for speaker i at frame $t - n$, and $R_{t-n}^{Evg(i)}$ denotes the enhanced sound source map obtained by weighted accumulation for speaker i. Weak sound sources may be overwhelmed by other high-energy sources or noise in a single frame. However, by leveraging weighted accumulation over multiple frames, their energy is aggregated in both spatial and temporal dimensions, whereas incoherent random noise is partially cancelled out, thereby improving the accuracy of weak sound source detection. The set of enhanced sound source maps for all speakers at frame t is defined as follows:

$$M_t = \left\{R_t^{Evg(i)}\right\}_{i=1}^{N_v}. \tag{8}$$

After the vision-guided weak sound source weighted enhancement method, all audio cues are further fed into the audio encoder to extract deep audio features. This process can be formulated as follows:

$$S_t = f_a^{enc}(M_t), S_t^G = f_a^{enc}\left(R_t^{gl}\right), \tag{9}$$

where $S_t = \left\{S_t^i\right\}_{i=1}^{N_v}$ represents the set of encoded audio features, and S_t^i denotes the audio feature of speaker i at frame t. S_t^G represents the encoded global audio feature.

3.2 | Audio-Visual Contrastive Learning

After obtaining the deep audio and visual representations through the measurement modules, it is crucial to align and integrate the multimodal features for consistent speaker representation. To this end, we introduce an audio-visual contrastive learning strategy, which not only enhances the consistency between modalities but also improves the discriminability across different speakers. The details are described as follows.

Contrastive learning is a self-supervised learning method that learns feature representations by pulling the features of positive sample pairs closer while pushing the features of negative sample pairs apart. The audio-visual contrastive learning (AVCL) mechanism is applied in the proposed AVCLNet to integrate visual and auditory modality representations into a unified feature space. By maximising the audio and visual feature similarity of the same target through pulling the features of audio-visual positive sample pairs (I_t, S_t) closer, the network is promoted to learn multimodal consistent representations for the same target, enhancing the correlation between audio and visual features. By pulling the features of visual positive sample pairs $\left(I_t, I_t^+\right)$ closer, the similarity of the same target across consecutive frames is maximised, improving the stability of visual measurement. I_t^+ represents the visual positive sample feature set, denoted as $I_t^+ = \left\{I_t^{i+}\right\}_{i=1}^{N_v}, I_t^{i+} = I_{t\pm n}^i, n \in \{1, 2, ..., n^+\}$. For each image frame V_t, we collect face detections of all speakers from the neighbouring frames within $\pm n$, forming a visual positive sample pool for each speaker. Then, one sample is randomly selected from each pool as the visual positive sample for the corresponding speaker and encoded as a visual positive sample feature I_t^{i+}. Taking $N_v = 3$ and $n^+ = 3$ as examples, the selection process of visual positive sample features is illustrated in Figure 4.

For the features of audio-visual positive sample pairs (I_t, S_t), a cross-modal contrastive loss $\mathcal{L}_{cc}$ is designed to enhance the correlation between target modalities, bringing the anchor speaker's visual features closer to the corresponding audio features while suppressing the interference from other speakers' visual and audio features. The definition is as follows:

$$\mathcal{L}_{cc} = -\frac{1}{2} \sum_{i=1}^{N_v} \left(\log \frac{\exp\left(\frac{I_t^i S_t^i}{\tau}\right)}{\sum_{\substack{j=1 \\ j \neq i}}^{N_v} \exp\left(\frac{I_t^i S_t^j}{\tau}\right)} + \log \frac{\exp\left(\frac{I_t^i S_t^i}{\tau}\right)}{\sum_{\substack{j=1 \\ j \neq i}}^{N_v} \exp\left(\frac{I_t^j S_t^i}{\tau}\right)} \right), \tag{10}$$

where the numerator represents the similarity between the audio and visual features of the anchor speaker. The denominator of the first term represents the sum of similarities between the anchor speaker's visual features and the audio features of other speakers, whereas the denominator of the second term represents the sum of similarities between the anchor speaker's audio features and the visual features of other speakers. τ is a learnable temperature parameter.

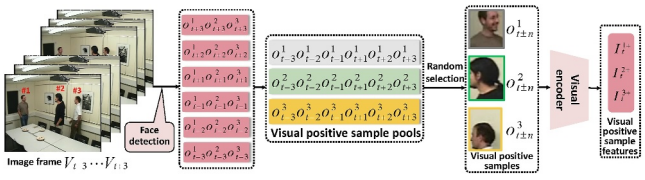

FIGURE 4 | Selection process of visual positive sample features: extract face regions from consecutive image frames, build visual positive sample pools, then randomly select samples from these pools and encode them as visual positive sample features.

For the features of visual positive sample pairs (I_t, I_t^+), an intra-visual-modal contrastive loss $\mathcal{L}_{vc}$ is designed to enhance the temporal consistency of the target, making the visual features of the anchor speaker more similar across consecutive frames while distinguishing the visual features of other speakers. The loss is defined as follows:

$$\mathcal{L}_{vc} = -\frac{1}{2}\sum_{i=1}^{N_v}\left(\log\frac{\exp\left(\frac{I_t^i I_t^{i+}}{\tau}\right)}{\sum_{\substack{j=1\\ j\neq i}}^{N_v}\exp\left(\frac{I_t^i I_t^j}{\tau}\right)}\right), \tag{11}$$

where the numerator represents the similarity of the anchor speaker's visual features across consecutive frames, whereas the denominator denotes the sum of similarities between the anchor speaker's visual features and those of other speakers. τ is a learnable temperature parameter.

Meanwhile, by pushing the features of negative sample pairs (I_t, I_t^-) and (S_t, S_t^-) apart, the similarity of different speakers' features is minimised, suppressing irrelevant visual information and audio signals while focusing on target-relevant visual and audio features to improve the accuracy of target association. The features of negative samples include the visual negative sample feature set I_t^- and the audio negative sample feature set S_t^-, denoted as $I_t^- = \{I_t^{i-}\}_{i=1}^{N_v}$, $I_t^{i-} = I_t^j$, $S_t^- = \{S_t^{i-}\}_{i=1}^{N_v}$, $S_t^{i-} = S_t^j$, where $i, j \in \{1, 2, ..., N_v\}$, $i \neq j$ representing the speaker ID. Audio and visual negative samples are randomly selected from the audio and visual negative sample pools of speakers different from the anchor speaker. Taking $N_v = 3$ as an example, the selection process of audio and visual negative sample features is illustrated in Figure 5.

For all the features of sample pairs $(I_t^i, S_t^i, I_t^{i+}, I_t^{i-}, S_t^{i-})$, the audio-visual quintuplet loss $\mathcal{L}_{qui}$ is designed to ensure that positive sample features are closer to the anchor speaker's features than negative sample features through margin constraints. It is defined as follows:

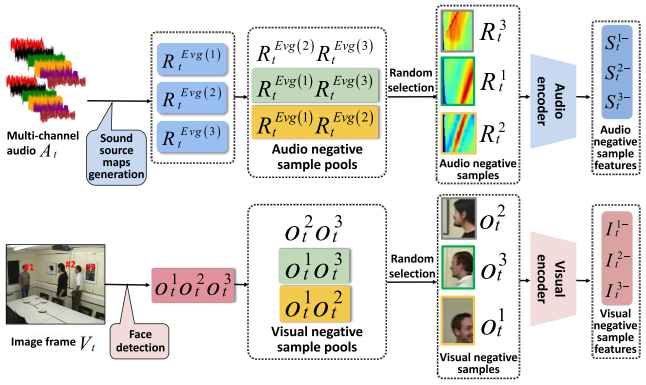

FIGURE 5 | Selection process of audio and visual negative sample features: sound source maps and face regions of nonanchor speakers are extracted to build negative sample pools, from which samples are randomly selected and encoded as audio and visual negative sample features.

$$\mathcal{L}_{qui} = \sum_{i=1}^{N_v}\sum_{j=1,j\neq i}^{N_v}\max\Big(\max\left\|I_t^i - S_t^i\right\| + \max\left\|I_t^i - I_t^{i+}\right\| - \min\left\|I_t^i - I_t^{i-}\right\| - \min\left\|S_t^i - S_t^{i-}\right\| + \varepsilon, 0\Big), \tag{12}$$

where $\|\cdot\|$ denotes the cosine distance between two embeddings, whereas ε represents the margin. Therefore, the total loss of AVCLNet is defined as follows:

$$\mathcal{L}_{Total} = \lambda\mathcal{L}_{cc} + \mu\mathcal{L}_{vc} + \eta\mathcal{L}_{qui}, \tag{13}$$

where the weights λ, μ and η are, respectively, assigned to the losses $\mathcal{L}_{cc}$, $\mathcal{L}_{vc}$ and $\mathcal{L}_{qui}$, to balance the contribution of each loss in the total loss function.

The audio and visual features obtained after the AVCL mechanism are fed into the cross-modal attention module for audio-visual fusion:

$$\begin{aligned} F_t = {} & LN(I_t + MHA(Q_a, K_v, V_v)) \\ & \oplus LN(S_t + MHA(Q_v, K_a, V_a)), \end{aligned} \tag{14}$$

where $F_t = \{F_t^i\}_{i=1}^{N_v}$ represents the set of audio-visual fusion features. Then, F_t is combined with the global audio feature S_t^G to obtain the global fusion feature $\hat{F}_t$, addressing complex scenarios such as missed visual detections. Finally, the global fused features are fed into the prediction head to obtain the detection results $\hat{p}_t = \{B_t^1, B_t^2, ..., B_t^{N_t}\}$.

3.3 | Data Association With Dual Geometric Constraints

To enhance the robustness of data association under different target motion patterns, we propose a dual geometric constraint strategy. By integrating the 2D bounding box IoU and 3D Euclidean distance constraints to construct the cost matrix, our method simultaneously accounts for spatial overlap and motion continuity, outperforming single-constraint approaches. Firstly, detection box matching serves as the foundation of data association. Given the detected bounding boxes B_{t-1}^i and B_t^j in adjacent frames, their intersection over union (IoU) is computed as the 2D cost:

$$C_{i,j}^{2D} = 1 - \frac{Area\left(B_{t-1}^i \cap B_t^j\right)}{Area\left(B_{t-1}^i \cup B_t^j\right)}, \tag{15}$$

where B_{t-1}^i and B_t^j, respectively, represent the bounding boxes of targets with ID = i and ID = j in frame $t-1$ and frame t. Through the camera model, the 3D centre coordinates of the target bounding box (a, b, c) are obtained, and the 3D Euclidean distance of the target across consecutive frames is defined as follows:

$$D_{i,j}^{3D} = \sqrt{\Delta a^2 + \Delta b^2 + \Delta c^2}, \Delta a = a_t^j - a_{t-1}^i. \tag{16}$$

To eliminate scale discrepancies, the similarity score metric is mapped using a Gaussian kernel function as the 3D cost:

$$C_{i,j}^{3D} = 1 - \exp\left(- \frac{\left(D_{i,j}^{3D}\right)^2}{2\sigma^2} \right), \tag{17}$$

where σ is dynamically adjusted based on the scene depth range to enhance the scene adaptability of distance sensitivity. The final matching cost C^m is obtained by fusing the two aforementioned costs through the cost weight β:

$$C_{i,j}^{m} = \beta \cdot C_{i,j}^{2D} + (1 - \beta) \cdot C_{i,j}^{3D}. \tag{18}$$

Our data association module uses the aforementioned constraints to associate trajectories and detections between adjacent frames. Specifically, the cost matrix between adjacent frames is computed as follows:

$$C = \begin{bmatrix} C_{1,1}^{m} & \cdots & C_{1,N_t}^{m} \\ C_{2,1}^{m} & \cdots & C_{2,N_t}^{m} \\ \vdots & \ddots & \vdots \\ C_{N_{t-1},1}^{m} & \cdots & C_{N_{t-1},N_t}^{m} \end{bmatrix}, \tag{19}$$

where N_{t-1} and N_t represent the number of tracked targets in the previous frame $t-1$ and the number of detected targets in the current frame t. $C_{i,j}^{m}$ denotes the matching cost between the tracking trajectory of the target with ID $= i$ in the previous frame and the detection result of the target with ID $= j$ in the current frame. Based on the above cost matrix, we use a greedy algorithm to determine the association between the trajectory before frame t (Tr_{t-1}) and the detection results at frame t ($\hat{p}_t = \{B_t^1, B_t^2, ..., B_t^{N_t}\}$). The core idea of the greedy algorithm is to select the trajectory-detection pair with the minimum matching cost at each step and add the current detection result to the previous trajectory until all detection results are matched and the trajectory Tr_t is updated.

4 | Experiments and Discussions

4.1 | Experimental Settings

To comprehensively evaluate the performance of AVCLNet, we conduct extensive experiments under standardised protocols. This section first introduces the datasets used for training and validation. We then define the evaluation metrics aligned with the objectives of the tracking task, ensuring fair comparisons with existing methods. Finally, the implementation details are elaborated, covering network configurations, hyperparameter settings and hardware environments.

4.1.1 | Dataset

AV16.3 [60] is an audio-visual corpus widely used to evaluate speaker localisation and tracking systems. The audio data (16 channels) are recorded at a 16-kHz sampling rate by two circular eight-element microphone arrays mounted on the table, spaced 0.8 m apart. The video data (288 × 360 pixels) are captured at a frequency of 25 Hz by monocular colour cameras installed at three corners of the room. In the experiments, two microphone arrays and one of the three cameras are selected for recording to evaluate the algorithm's performance under different viewpoints. In the sequences, 2–3 participants speak and engage in various activities in a conference room, including sitting statically, standing statically or walking around the table. Each sequence lasts approximately 20–60 s. We train the model on 13,450 audio-visual sample pairs from the multispeaker sequences *seq*18, 19, 35, 40 and evaluate it on *seq*24, 25, 30, 45.

CAV3D [61] is an audio-visual speaker tracking corpus collected by a co-located sensor platform. The dataset is collected in a room of size ($4.77m \times 5.95m \times 4.5m$). The audio data (8 channels) are recorded at a 96-kHz sampling rate by a circular eight-element microphone array. The video data (768 × 1024 pixels) are captured at a frequency of 15 Hz by a camera with a 90° field of view. This dataset contains 5 multispeaker sequences. Compared to AV16.3, the scenes in the CAV3D dataset are more challenging, including scenarios with 2–3 speakers involving mutual occlusion, entering or exiting the camera's field of view and periods of silence. Each sequence lasts approximately 60–90 s. We select data from 3 sequences in the CAV3D-MOT (multispeaker sequences), with a total of 11,935 audio-visual sample pairs used for model training, and test the model using data from the remaining 2 sequences.

Both datasets offer synchronised and well-calibrated multimodal recordings, featuring temporally aligned audio-visual streams, precise camera calibration parameters and comprehensive identity-level annotations. These characteristics are critical for effectively evaluating cross-modal fusion strategies and speaker identity association mechanisms. Accordingly, AV16.3 and CAV3D are selected as the primary evaluation benchmarks for this study.

4.1.2 | Metrics

Mean absolute error (MAE) is a metric used to evaluate tracking performance. It measures the accuracy of target localisation by calculating the Euclidean distance between the predicted position and the ground truth position. Because MAE directly reflects the magnitude of localisation error, it is widely used to compare the performance of different algorithms. The definition of MAE is as follows:

$$MAE = \frac{1}{N_t T} \sum_{i=1}^{N_t} \sum_{t=1}^{T} \left\| \hat{p}_{t,i} - \hat{p}_{t,i}^{gt} \right\|_2, \tag{20}$$

where N_t denotes the number of targets in each frame, T is the total number of frames, and $\hat{p}_{t,i}$ and $\hat{p}_{t,i}^{gt}$, respectively, represent the predicted position and ground truth position.

Multiple object tracking accuracy (MOTA) reflects the success rate of the tracker in detecting targets and associating trajectories. It is used to measure various errors that occur during tracking, including IDS (identity switch), FP (detection error exceeding the predefined threshold) and FN (missed targets). We record results exceeding 1/15 of the image diagonal size as FP. The definition of MOTA is as follows:

$$MOTA = \left(1 - \frac{\sum_t (IDS_t + FP_t + FN_t)}{\sum_t N_t^{gt}}\right) \times 100, \tag{21}$$

where N_t^{gt} denotes the number of ground truth targets in frame t.

Multiple object tracking precision (MOTP) reflects the accuracy of correctly tracked target positions and is used to measure the spatial precision of the tracker, that is, the average error between the predicted positions and the ground truth positions. The definition of MOTP is as follows:

$$MOTP = \frac{\sum_{i,t} e_t^i}{\sum_{i,t} m_t}, \tag{22}$$

where e_t^i denotes the Euclidean distance between the predicted target with ID = i and its matched ground truth, and m_t represents the number of successful matches at time step t.

4.1.3 | Implementation Details

In visual measurement, the face detection module adopts a YOLOv10 model [62] pretrained on the ImageNet dataset [63]. In audio measurement, the speech signals from the circular microphone array undergo a 40ms framing process, with Hamming windowing applied and a frame shift of 1/2 frame length. The interval between 2D sampling points is 3 pixels. Based on the actual room configurations of the two datasets, 3D sampling points located outside the room boundaries and below the table surface are removed. Using the camera calibration parameters provided by the dataset, a pinhole camera model is established to achieve 3D image projection. When calculating the weak sound source enhancement weight, we set $n = 3$. For the computation of the total loss $\mathcal{L}_{Total}$, we set $\tau = 0.1$, $\lambda = 0.4$, $\mu = 0.3$ and $\eta = 0.3$. For the computation of the cost matrix, we set $\sigma = 0.5m$ and $\beta = 0.6$. The model is optimised using the SGD optimiser with a learning rate of 1×10^{-4}. The model is trained for 50 epochs with a batch size of 16. The experiments are conducted on the PyTorch framework with one NVIDIA RTX 4090 Ti GPU.

4.2 | Comparisons With State-Of-The-Art Methods

Our approach is compared with unimodal methods and previous state-of-the-art audio-visual methods on the AV16.3 and CAV3D datasets, with the results presented in Tables 1 and 2. The audio-only (AO) and visual-only (VO) methods are implemented based on the audio and visual measurements described in Section 3.1. Unimodal methods rely solely on a single information source. Evidently, compared to unimodal approaches, the fusion of audio and visual modalities provides significant advantages in speaker tracking tasks. We reproduce the methods in Refs. [12, 22, 61] by running their publicly available source codes, whereas the results of Ref. [64] are directly referenced from the published work. As shown in Table 1, our AVCLNet outperforms all baseline methods across all evaluation metrics. Specifically, the MAE and MOTP are reduced to 10.70 pixels and 5.65 pixels, respectively, whereas MOTA is

TABLE 1 | Experimental results of unimodal methods and state-of-the-art audio-visual methods on the AV16.3 dataset.

Sequences		MAE ↓						MOTA ↑						MOTP ↓					
Seq	cam	AO	VO	[22]	[61]	[14]*	[12]	Ours	AO	VO	[22]	[61]	[14]*	[64]	[12]	Ours	AO	VO	[22]	[61]	[14]*	[12]	Ours
24	1	48.73	27.32	18.43	4.45	20.94	4.26	13.38	51.56	65.51	78.57	97.25	76.51	89.40	97.85	81.59	12.82	8.65	10.48	4.42	7.42	4.16	7.60
	2	40.25	29.66	12.22	38.21	6.91	5.02	4.95	61.69	64.17	79.58	61.02	85.28	91.00	94.16	96.63	9.46	9.48	9.11	6.09	5.16	4.24	4.04
	3	44.10	32.21	14.75	34.55	5.83	25.49	15.79	55.43	59.40	59.46	57.71	86.33	91.00	61.36	73.19	10.80	9.98	10.78	8.69	4.62	4.93	4.58
25	1	52.64	21.30	15.55	12.33	5.14	20.79	8.70	48.31	85.71	82.18	70.57	86.42	94.80	64.91	96.01	14.21	7.12	14.32	8.69	4.57	6.25	4.12
	2	44.89	27.89	19.12	9.12	7.01	6.51	9.56	54.86	72.20	64.44	84.37	84.02	85.90	90.49	86.87	10.23	7.99	8.14	7.94	5.18	5.13	6.87
	3	43.97	35.34	12.40	9.15	9.23	10.16	12.21	56.17	56.82	80.62	87.55	82.86	91.20	82.72	80.36	11.72	11.28	11.30	6.85	5.49	6.11	7.05
30	1	48.37	48.91	16.91	7.23	18.32	16.33	10.65	50.88	52.89	69.85	88.35	77.21	40.80	70.45	82.58	13.40	12.90	12.48	6.92	6.63	6.18	6.14
	2	76.52	35.43	11.42	6.69	19.68	4.85	3.98	41.59	61.85	86.67	97.15	75.42	82.70	99.53	98.22	18.61	9.22	10.49	6.92	6.97	4.63	3.59
	3	59.26	30.96	11.30	5.16	14.89	4.72	11.35	46.72	67.56	58.41	96.50	78.27	64.70	99.42	78.24	15.64	8.09	13.01	5.12	6.13	4.72	4.77
45	1	53.64	32.26	20.04	16.35	22.84	17.30	9.14	48.98	61.47	63.14	69.58	73.94	87.50	75.57	89.52	14.09	10.01	16.59	12.66	8.01	8.94	4.25
	2	83.13	50.11	24.64	25.13	21.31	16.45	18.56	38.06	47.00	62.42	60.33	74.58	88.30	85.90	66.43	19.36	12.57	22.20	9.74	7.89	7.02	7.97
	3	77.35	39.73	22.97	19.17	20.17	16.03	10.27	41.21	58.25	68.97	70.02	75.19	80.10	79.82	81.69	17.86	9.98	19.46	7.90	7.62	7.59	6.81
Average		56.07	34.34	16.65	15.63	14.36	12.32	**10.70**	49.62	62.73	71.19	78.36	83.05	77.92	83.18	**83.94**	14.01	9.77	13.19	7.46	6.31	5.83	**5.65**

Note: Bold values indicate the best performance.

TABLE 2 | Experimental results of unimodal methods and state-of-the-art audio-visual methods on the CAV3D dataset.

Sequences	2D	AO	VO	[14]*	[61]	[12]	Ours
MOT	MAE↓	39.61	22.19	15.02	10.10	12.38	**10.04**
	MAE∗ ↓	9.55	5.67	5.14	4.90	4.82	**4.56**

Note: Bold values indicate the best performance.

improved to 83.94%, demonstrating the superiority of AVCLNet in multispeaker tracking, particularly in dynamic multispeaker (three speakers) interaction scenarios. The sequential Monte Carlo-probability hypothesis density (SMC-PHD) filtering-based method [22] incurs high computational costs in multitarget scenarios, making real-time processing infeasible. Additionally, its sparse sampling strategy leads to degraded tracking accuracy in dense target and dynamic interaction scenes. We replaced the original feature encoder in AVCLNet with the pretrained AV-HuBERT [14] model to extract audio-visual features while retaining our proposed data association modules. Experimental results show that the AV-HuBERT-based variant [14]* performs worse than AVCLNet across all evaluation metrics. This is mainly because AV-HuBERT [14] is designed for token-level semantic alignment between audio and visual streams (e.g., voice-lip correspondence), but it lacks spatial modelling capability and identity-level discrimination, which are crucial for multispeaker tracking tasks. Methods in Refs. [12, 61] lack explicit identity consistency optimisation, resulting in a high ID-switch rate and failing to address the problem of identity ambiguity. In addition, these methods are unable to effectively handle the issue of weak sound sources being easily masked by noise in multitarget scenarios. In contrast, our AVCLNet integrates contrastive learning strategies with spatiotemporal modelling capabilities. By employing identity-aware contrastive learning, we construct a unified feature space that effectively mitigates feature bias caused by cross-modal heterogeneity, significantly reducing frequent identity switches. Meanwhile, the vision-guided temporal weighted strategy enhances weak sound source localisation under noisy conditions, and the dual geometric constraint strategy jointly considers spatial overlap and motion continuity, reducing identity switch and matching errors.

On the more challenging CAV3D dataset, AVCLNet also demonstrates superior performance, as shown in Table 2. We introduce MAE∗ following Ref. [61], which represents the MAE computed only on successfully tracked frames. AVCLNet achieves the MAE∗ as low as 4.56 pixels, outperforming other audio-visual trackers and indicating higher localisation accuracy in stable tracking scenarios.

4.3 | Ablation Study

4.3.1 | The Effectiveness of Each Component

To validate the effectiveness of each module, we conduct ablation experiments on the visual-guided weak sound source weighted enhancement (V-WSWE) module, the audio-visual contrastive learning (AVCL) module and the dual geometric constraint (DGC) module. The results are shown in Table 3 and visualised in Figure 6. After removing the V-WSWE module (denoted as w/o V-WSWE), the MAE on the AV16.3 and CAV3D datasets, respectively, increases from 10.70 pixels and 10.04 pixels in the baseline model to 13.23 pixels (an increase of 2.53 pixels) and 14.08 pixels (an increase of 4.04 pixels). This result indicates that the visual-guided cross-modal mapping, by utilising face detection results to constrain the sound source sampling range, effectively suppresses noise interference. The temporal dynamic weighted method, through the weighted accumulation of acoustic clues from historical frames, enhances the detectability of weak sound sources in both spatial and temporal dimensions. Further removing the AVCL module (denoted as w/o V-WSWE + AVCL), the MAE, respectively, increases to 29.53 pixels (an increase of 16.30 pixels) and 29.99 pixels (an increase of 15.91 pixels) on the two datasets, whereas MOTA sharply drops to 65.10% and 65.05%. This result highlights that the contrastive learning mechanism effectively addresses the feature bias caused by modality heterogeneity, improving tracking accuracy. By pulling close audio and visual features of the same target while pushing apart those of different targets, it enhances the consistency of identity representation. Finally, further removing the DGC module (denoted as w/o V-WSWE + AVCL + DGC), MOTA decreases significantly, from 65.10% to 55.93% on the AV16.3 dataset and from 65.05% to 52.34% on the CAV3D dataset. This shows that the 3D geometric constraint, by calculating the Euclidean distance between targets in consecutive frames, compensates for the limitations of relying solely on 2D bounding box IoU, significantly improving data association robustness in complex motion scenarios. The combination of 2D and 3D geometric information reduces mismatches caused by viewpoint changes or target occlusions. The experimental results demonstrate that the collaboration of each module jointly supports the efficiency and robustness of AVCLNet in multispeaker tracking tasks. The experimental results indicate that the collaborative effect of all modules jointly supports the efficiency and robustness of AVCLNet in multispeaker tracking tasks.

4.3.2 | Visual-Guided Weak Sound Source Weighted Enhancement Module

To validate the effectiveness of the proposed audio measurement module, we conduct an ablation study based on the audio measurement module and compare the following three variants. Firstly, the authors in Ref. [11] proposed the spatiotemporal GCF (stGCF) method, which incorporates spatial and temporal information assisted by the visual modality to improve sound source localisation accuracy. However, due to the lack of robust modelling for interference and noise in multispeaker scenarios, stGCF suffers significant performance degradation in complex environments. Secondly, the authors in Ref. [12] introduced a visual-guided GCF (vgGCF) method, which establishes a mapping between audio and visual representations to achieve the

TABLE 3 | Ablation study results of each component in the AVCLNet.

Module	MAE ↓		MOTA ↑		MOTP ↓	
	AV16.3	CAV3D	AV16.3	CAV3D	AV16.3	CAV3D
AVCLNet	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**
W/o V-WSWE	13.23	14.08	79.48	76.64	6.97	7.41
W/o V-WSWE + AVCL	29.53	29.99	65.10	65.05	10.32	10.86
W/o V-WSWE + AVCL + DGC	32.44	33.60	55.93	52.34	11.85	12.19

Note: Bold values indicate the best performance.

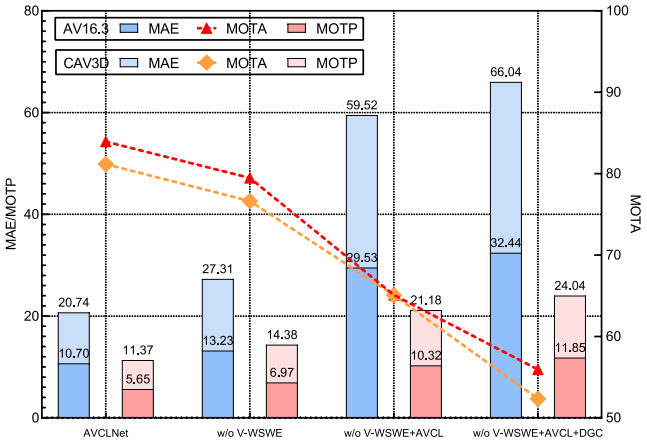

FIGURE 6 | Ablation study results of each component in the AVCLNet on AV16.3 and CAV3D datasets.

fusion of heterogeneous modalities within a unified localisation space. Although vgGCF integrates visual cues, it fails to differentiate between sound sources with varying intensities, making it difficult to distinguish weak or overlapping sound sources effectively. In contrast, our proposed visual-guided weak sound source weighted enhancement (V-WSWE) method introduces a visual-guided weighted mechanism that not only accurately focuses on salient sound sources but also significantly enhances the discrimination of weak sound regions. This leads to more robust and precise performance in multispeaker tracking tasks. As shown in Table 4 and visualised in Figure 7, V-WSWE significantly outperforms the other two methods across all evaluation metrics on both the AV16.3 and CAV3D datasets. Specifically, V-WSWE achieves the lowest MAE (34.34 pixels and 38.69 pixels), the highest MOTA (62.73% and 60.03%) and superior localisation precision in terms of MOTP (9.77 pixels and 10.10 pixels). These results demonstrate the clear advantage of our visual-guided weighted mechanism under complex multispeaker scenarios.

4.3.3 | Audio-Visual Contrastive Learning

We introduce multiple loss functions in the module of audio-visual contrastive learning, including cross-modal contrastive loss $\mathcal{L}_{cc}$, intra-visual-modal contrastive loss $\mathcal{L}_{vc}$ and audio-visual quintuplet loss $\mathcal{L}_{qui}$. To validate the contribution of each loss function to the model's performance, we conduct an ablation study. The experimental results are shown in Table 5, demonstrating that the model performs optimally when all loss functions are used together. The audio-visual quintuplet loss $\mathcal{L}_{qui}$ enhances the discriminability of cross-modal features by optimising the cosine distance between positive and negative sample pairs through margin constraints, thereby suppressing identity switch. After removing $\mathcal{L}_{qui}$, the decrease in the tracker's ability to distinguish target features leads to an increase in identity switch rate, with MOTA on the two datasets decreasing from 83.94% to 81.77% and from 81.17% to 79.98%, confirming the key role of this loss in identity matching for multitarget tracking. The intra-visual-modal contrastive loss $\mathcal{L}_{vc}$ strengthens the consistency of visual features of the same target across time, reducing tracking drift caused by occlusion or abrupt viewpoint changes. After removing $\mathcal{L}_{vc}$, the tracker's temporal association ability significantly deteriorates, with MAE, respectively, increasing to 12.91 pixels and 12.46 pixels on the two datasets, whereas MOTA decreases to 80.23% and 77.73%, indicating that this loss enhances the robustness of temporal association. The cross-modal contrastive loss $\mathcal{L}_{cc}$ aligns the visual and audio features into a unified space, addressing the representation bias caused by modality heterogeneity. After removing $\mathcal{L}_{cc}$, the cross-modal feature bias causes a significant drop in tracking accuracy, with MAE, respectively, increasing to 22.37 pixels and 22.25 pixels on the two datasets, whereas MOTA decreases to 71.10% and 71.54%, confirming the core contribution of this loss to feature alignment. The experiments show that the multiloss collaborative optimisation framework in audio-visual contrastive learning effectively balances cross-modal alignment, temporal consistency and identity discrimination, providing comprehensive support for multispeaker tracking in complex scenarios.

4.3.4 | Dual Geometric Constraint Module

To further validate the effectiveness of the proposed dual geometric constraints (DGC) module in the data association phase, we design an ablation study to compare and analyse the performance changes after removing the 2D or 3D geometric constraints. The results are shown in Table 6. In the complete DGC module, both the 2D bounding box IoU constraint and the 3D Euclidean distance constraint are introduced, implementing dual geometric constraints at the image and spatial levels, thereby effectively improving target matching accuracy and trajectory consistency. After removing the 2D geometric constraint (DGC (w/o 2D)) and the 3D geometric constraint (DGC (w/o 3D)), the model shows a noticeable decline in all performance metrics. Numerically, the complete DGC module achieves the lowest MAE of 10.70 pixels and 10.04 pixels, the highest MOTA of 83.94% and 81.17% and the best MOTP of 5.65 pixels and 5.72 pixels on the AV16.3 and CAV3D datasets,

TABLE 4 | Ablation study results of audio measurement using a visual-guided weak sound source weighted enhancement module.

Method	MAE ↓		MOTA ↑		MOTP ↓	
	AV16.3	CAV3D	AV16.3	CAV3D	AV16.3	CAV3D
stGCF	61.21	63.82	46.78	44.53	13.53	13.99
vgGCF	41.35	40.86	56.54	52.90	9.85	11.34
V-WSWE	**34.34**	**38.69**	**62.73**	**60.03**	**9.77**	**10.10**

Note: Bold values indicate the best performance.

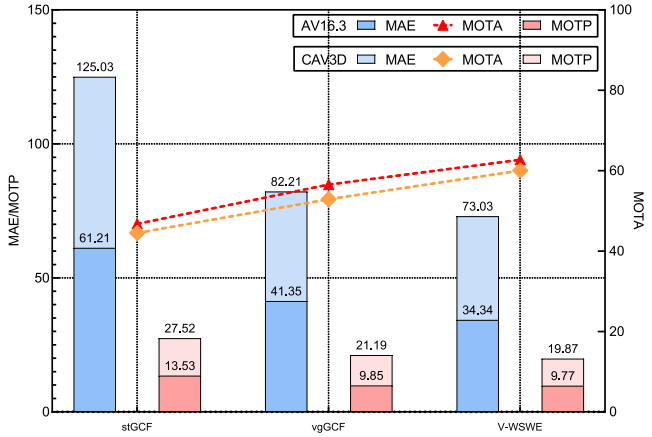

FIGURE 7 | Ablation study results of the visual-guided weak sound source weighted enhancement module in the form of line and bar chart.

respectively. More notably, in the identity switches (IDS) metric, the DGC module achieves the fewest identity switch occurrences (147 and 163), significantly outperforming the other two variants. The IDS metric represents the number of times the same target is incorrectly assigned different IDs across different frames in multiobject tracking, reflecting the consistency and robustness of the model in maintaining target identity during the tracking and data association process. The lower the IDS, the more stable and accurate the tracker is at maintaining target identity, which is particularly crucial in multispeaker scenarios. From the comparison, it can be observed that when the 2D geometric constraint is removed, the model shows a certain degree of performance degradation in both MOTA and IDS metrics (e.g., MOTA in AV16.3 drops from 83.94% to 78.25%, and IDS increases from 147 to 166). When the 3D geometric constraint is removed, the performance degradation is even more significant, with IDS rising to 211 and 234. This indicates that multiple constraints can compensate for the information loss caused by a single constraint, allowing for a more accurate capture of the target's appearance features and spatial position, thereby effectively reducing IDs and improving the accuracy of data association.

4.4 | Effect of Key Hyperparameters

We conducted sensitivity analysis on several key hyperparameters of the model, as shown in Table 7, including the temperature parameter τ, the visual positive sample selection range n^+, and the balancing coefficients λ, μ and η in audio-visual contrastive learning in Section 3.2, as well as the cost weight β in data association in Section 3.3. These parameters were all tuned through experiments on the validation set and kept fixed in all main experiments.

The temperature parameter τ in contrastive learning controls the sharpness of the feature similarity distribution. We set $\tau = 0.1$. Experiments demonstrate that this value achieves a good balance between convergence speed and feature discriminability. A smaller value (e.g., $\tau = 0.05$) leads to slower training and a higher risk of overfitting, whereas a larger value (e.g., $\tau = 0.2$) results in insufficient feature clustering and degrades contrastive performance.

The selection range n^+ for visual positive samples represents the temporal sampling window size of the visual modality, used to construct positive sample pairs. We conducted sensitivity experiments with $n^+ \in \{1, 3, 5, 10\}$, and the results show that $n^+ = 3$ strikes a balance between temporal diversity and semantic consistency of visual features from the same speaker. When $n^+ = 1$, although the positive sample pairs have very high semantic consistency, the limited temporal span makes the contrastive task too simple and prone to local optima. When $n^+ = 5$ or $n^+ = 10$, the temporal span of the positive sample pairs increases, which helps learn more discriminative visual features but introduces more cross-event or cross-speaker samples, thus weakening the semantic consistency of positive sample pairs.

In the audio-visual contrastive learning module, to balance the contribution of each loss component to the total loss function, we set the weights of the cross-modal contrastive loss $\mathcal{L}cc$, intra-visual-modal contrastive loss $\mathcal{L}vc$ and audio-visual quintuplet loss $\mathcal{L}qui$ as $\lambda = 0.4$, $\mu = 0.3$ and $\eta = 0.3$, respectively. This configuration emphasises the importance of cross-modal representation alignment by assigning a higher weight to $\mathcal{L}cc$ to strengthen the consistency between visual and audio features. At the same time, appropriate constraints on temporal consistency in the visual modality ($\mathcal{L}vc$) and margin-based discrimination between audio-visual positive and negative samples ($\mathcal{L}qui$) are introduced to jointly enhance the discriminative capability and association accuracy of the network for multimodal targets. Multiple ablation studies confirm that this weight configuration achieves a good balance between performance and training stability.

The cost weight β is used to balance the relative importance of 2D geometric constraints and 3D geometric constraints in the data association strategy. We conducted a sensitivity analysis on $\beta \in \{0.4, 0.6, 0.8\}$. Experimental results show that the best performance is achieved when $\beta = 0.6$. Specifically, when $\beta = 0.4$, the model relies on 3D Euclidean distance, which is susceptible to depth estimation errors, leading to association failures, especially in weak sound source scenarios. When $\beta = 0.8$, the

TABLE 5 | Ablation study results of loss functions in audio-visual contrastive learning.

			MAE ↓		MOTA ↑		MOTP ↓	
$\mathcal{L}_{cc}$	$\mathcal{L}_{vc}$	$\mathcal{L}_{qui}$	AV16.3	CAV3D	AV16.3	CAV3D	AV16.3	CAV3D
✓	✓	✓	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**
✓	✓	—	12.45	12.87	81.77	79.98	6.43	7.23
✓	—	✓	12.91	12.46	80.23	77.73	6.17	6.89
—	✓	✓	22.37	22.25	71.10	71.54	9.08	10.02
—	—	✓	26.67	27.07	67.12	64.22	10.09	10.26

Note: Bold values indicate the best performance.

TABLE 6 | Ablation study results of each component in the dual geometric constraint module.

	MAE ↓		MOTA ↑		MOTP ↓		IDS ↑	
Module	**AV16.3**	**CAV3D**	**AV16.3**	**CAV3D**	**AV16.3**	**CAV3D**	**AV16.3**	**CAV3D**
DGC	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**	**147**	**163**
DGC (w/o 2D)	12.02	12.68	78.35	77.64	5.78	5.94	166	179
DGC (w/o 3D)	14.58	16.17	71.14	71.03	6.99	7.26	211	234

Note: Bold values indicate the best performance.

TABLE 7 | Impact of key hyperparameters on tracking performance on AV16.3 and CAV3D datasets. Each row varies one parameter (or loss weight combination) while keeping others fixed.

		MAE ↓		MOTA ↑		MOTP ↑	
Parameter	**Value**	**AV16.3**	**CAV3D**	**AV16.3**	**CAV3D**	**AV16.3**	**CAV3D**
τ	0.05	20.45	21.03	76.12	74.85	10.01	9.94
	0.10	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**
	0.20	17.92	18.41	78.85	77.32	8.13	9.02
n^+	1	23.81	25.34	73.85	72.49	10.13	10.57
	3	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**
	5	12.02	12.41	81.87	80.20	6.27	6.85
	10	15.72	16.96	78.61	76.93	8.95	8.66
(λ, μ, η)	(0.2, 0.4, 0.4)	20.26	21.14	75.44	73.29	9.80	10.33
	(0.4, 0.3, 0.3)	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**
	(0.6, 0.2, 0.2)	19.68	18.26	77.49	75.68	9.28	9.87
β	0.4	11.64	11.33	82.03	80.50	5.91	6.06
	0.6	**10.70**	**10.04**	**83.94**	**81.17**	**5.65**	**5.72**
	0.8	12.91	13.90	79.42	78.10	6.74	7.02

Note: Bold values indicate the best performance.

model relies on 2D bounding box IoU on the image plane, ignoring spatial differences in the real world, which easily causes ID switches between visually adjacent but spatially distant targets. Therefore, $\beta = 0.6$ achieves a good trade-off between spatial perception and image features, effectively improving the association stability and accuracy of the tracking system under multisource interference.

4.5 | Real-Time Performance and Computational Complexity Analysis

To validate the real-time performance and applicability of the proposed network, we conducted benchmark tests on the AVCLNet against representative state-of-the-art audio-visual multispeaker tracking methods on the AV16.3 validation set, including the AV-HuBERT variant [14]*, AV3T [61] and STNet [12]. All models were evaluated on the same NVIDIA RTX 3090 GPU. As shown in Table 8, AVCLNet achieved an inference speed of 36.74 FPS and a computational cost of 3.82 GFLOPs. Compared to the AV-HuBERT variant [14]*, the FLOPs were reduced by approximately 25.4%, as AV-HuBERT has a more complex model structure that includes a large number of transformer encoder layers for semantic modelling, leading to higher computational overhead and making it less suitable for real-time multispeaker tracking tasks. In contrast, AVCLNet is designed with an emphasis on efficient spatial modelling and identity discrimination, significantly reducing inference cost while maintaining

TABLE 8 | Real-time performance analysis of AVCLNet and representative audio-visual multispeaker tracking methods.

Method	[14]*	[61]	[12]	Ours (w/i stGCF)	Ours (w/i V-WSWE)
FPS	9.32	16.88	12.50	32.75	**36.74**
FLOPs (G)	20.78	6.47	5.12	4.25	**3.82**

Note: Bold values indicate the best performance.

performance. Compared to the particle-filter-based AV3T [61] method (with 100 particles), AVCLNet reduced the FLOPs by 81.6%. This is because particle filtering inherently relies on sampling and iterative updates of a large number of particles, and its inference overhead increases rapidly with the number of targets, making it difficult to meet the requirements of real-time deployment. AVCLNet, on the other hand, utilises a lightweight feature extraction network and a modular contrastive learning mechanism. Compared to the deep learning-based STNet [12], AVCLNet achieved approximately 118% higher FPS and 41.0% lower FLOPs. Although STNet can process multiple targets in parallel, it was originally designed for single-target tracking and lacks architectural optimisation for multispeaker scenarios. In multitarget cases, it suffers from high module redundancy, and its feature extraction and matching processes cannot be shared, resulting in low resource utilisation efficiency and limiting its inference speed and scalability. In contrast, AVCLNet is specifically designed for multispeaker scenarios with efficient modality fusion and data association mechanisms, making it more suitable for real-time multitarget tracking tasks.

We conducted a complexity analysis of the proposed AVCLNet framework and evaluated the computational load controllability of each submodule. The visual measurement module adopts a YOLO-based detector, whose complexity mainly stems from the multiscale convolutional structures in the backbone network and the prediction heads. The overall computational cost is $O(CHWq^2)$, where C denotes the number of channels, H and W represent the image resolution, and q^2 is the kernel size of the convolution. The complexity of the audio measurement module mainly lies in the time delay estimation and the generation of generalised cross-correlation (GCC) maps based on spatial sampling points. For each frame, k depth levels and p_{3d} spatial sampling points are projected into 3D space for acoustic delay estimation, resulting in a computational cost of approximately $O(kp_{3d}M)$, where M is the number of microphone pairs. In the audio-visual feature alignment stage, we design three types of contrastive loss functions. This module requires constructing positive and negative sample pairs and computing the cosine similarity or Euclidean distance between features. The overall complexity is $O\left(n_p^2 D\right)$, where n_p denotes the number of samples and D the feature dimensionality. In the data association stage, the dual geometric constraint (DGC) strategy combines the matching cost calculations of 2D IoU and 3D Euclidean distance. Its matching complexity is $O(N_t^2)$, where N_t is the number of detected targets in the current frame. To summarise, the core computational complexity of AVCLNet is mainly dominated by the following components: $O\left(CHWq^2 + kp_{3d}M + n_p^2 D + N_t^2\right)$. When the number of spatial sampling points is large, the audio measurement module becomes the dominant contributor to the overall complexity. To alleviate the computational burden at this stage, we further introduce a vision-guided weak sound source weighted enhancement (V-WSWE) strategy. This approach leverages visual detection results to provide a coarse localisation of sound sources, thereby restricting acoustic sampling to local regions in space and significantly reducing the number of p_{3d}. Experimental results show that this strategy substantially improves the overall inference efficiency.

4.6 | Visualisation Analysis

To demonstrate the role of the AVCL mechanism in enhancing audio-visual feature alignment and consistency, we present the visualisation results of different stages of AVCLNet on a 2-speaker sequence. As shown in Figure 8, Panel (a) shows the face detection results, and Panels (b) and (c) correspond to the audio-visual feature fusion heatmaps without AVCL (w/o AVCL) and with AVCL (w/i AVCL). From the comparison of these heatmaps, it is clear that the AVCL mechanism plays a crucial role in enhancing feature fusion. Specifically, in Panel (b), the response regions are more dispersed, especially in scenes with complex backgrounds or strong acoustic interference, making it prone to environmental noise. For instance, at time steps $t = 230$ and $t = 240$, the target region's heatmap responses shift, and some acoustic information is incorrectly mapped to nontarget areas. In contrast, Panel (c) exhibits more concentrated responses in the target area, thanks to the effect of the contrastive learning mechanism. By pulling the audio and visual feature representations of the same target closer and pushing the features of different targets apart, the cross-modal information aligns in a unified feature space, promoting audio and visual feature alignment and identity consistency. Similarly, at $t = 230$ and $t = 240$, Panel (c) shows a clear advantage, with its high-response areas strictly covering the target speaker's location and avoiding interference from environmental noise. Further analysis of the detection results, as shown in Panels (d) and (e), demonstrates that AVCL can accurately detect the target location compared to without AVCL and maintain consistency in detection results throughout the entire time sequence.

Figure 9 further presents the visualisation results on the more challenging CAV3D dataset. The selected sequence involves three simultaneously active speakers with frequent occlusions, complex background clutter and speaker motion. The visualisations show that AVCLNet maintains accurate and robust tracking performance, effectively handling identity consistency and cross-modal alignment in more dynamic and realistic scenarios.

4.7 | Limitations and Future Work

Although AVCLNet demonstrates strong cross-modal robustness in complex scenarios, it still has certain limitations. In

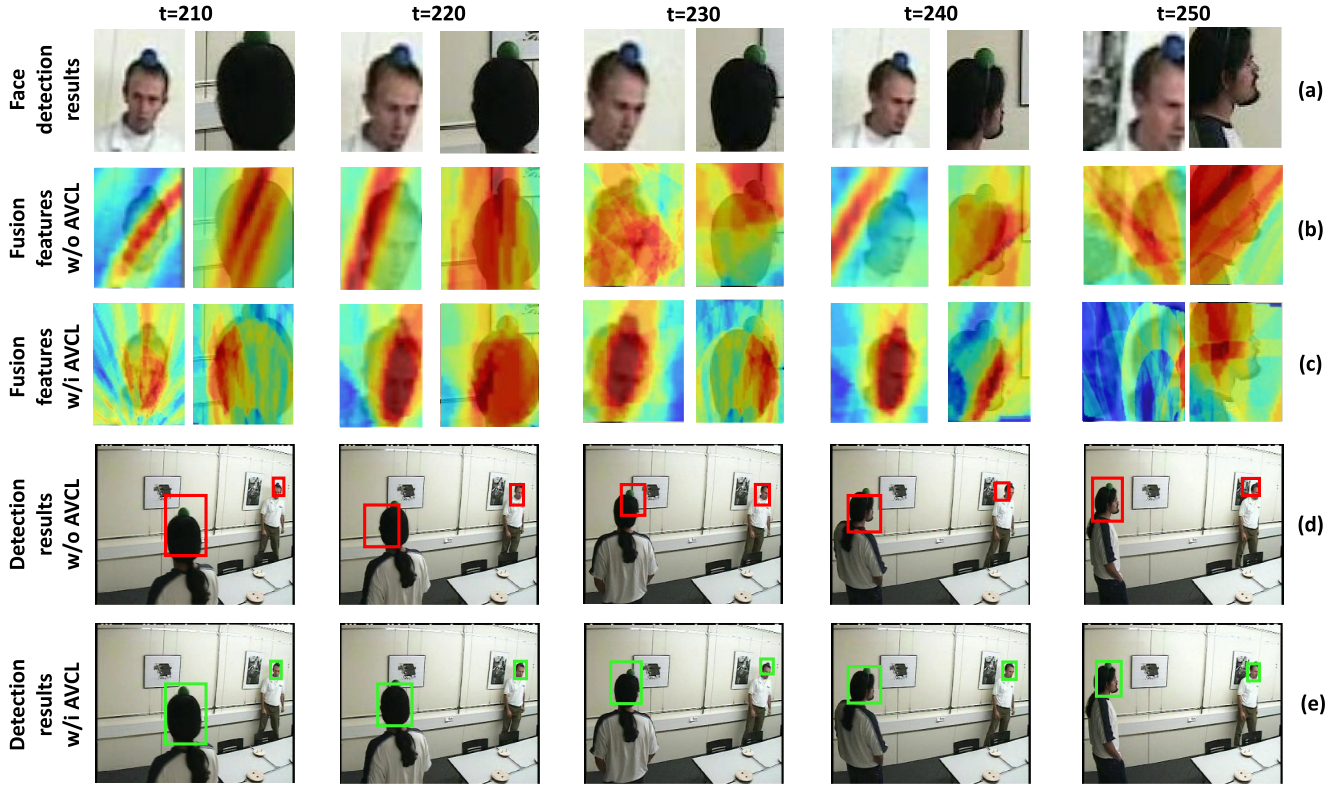

FIGURE 8 | Visualisation results on the AV16.3 dataset seq24-cam1 sequence. (a) Face detection results of visual measurement, (b) audio-visual feature fusion heatmap w/o AVCL, (c) audio-visual feature fusion heatmap w/i AVCL, (d) detection results w/o AVCL and (e) detection results w/i AVCL.

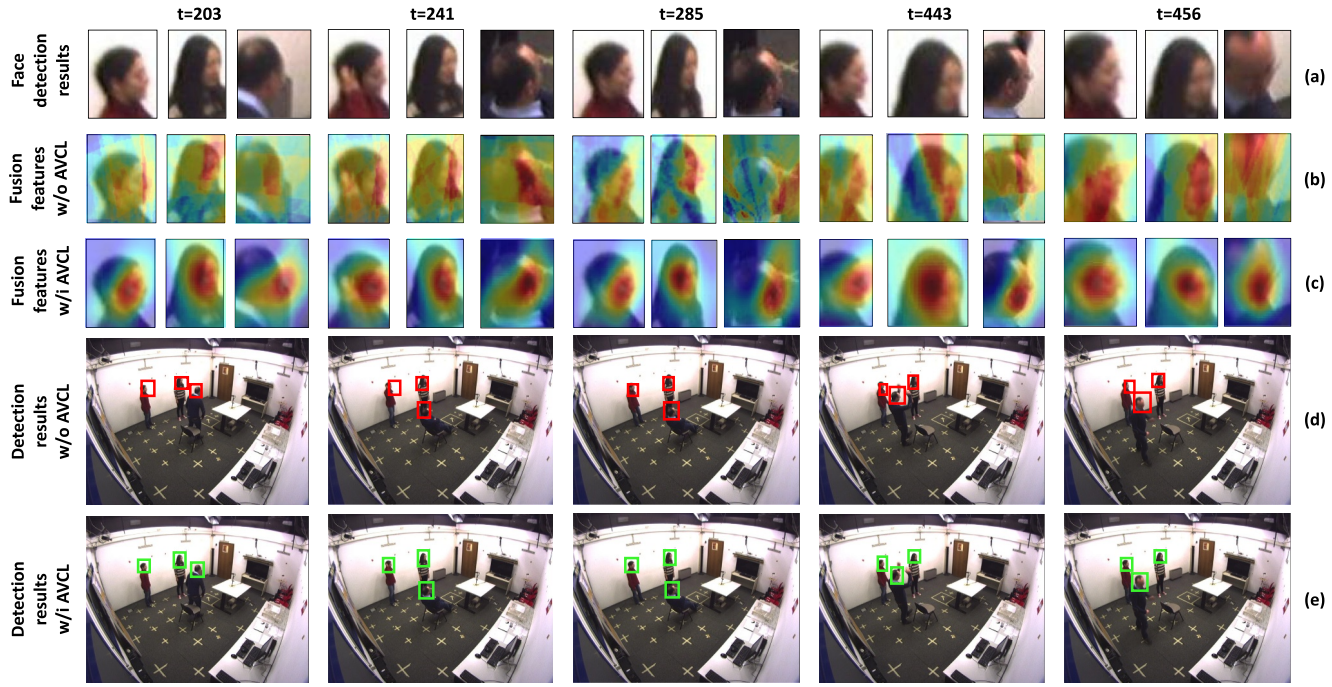

FIGURE 9 | Visualisation results on the CAV3D dataset seq26-cam4 sequence. (a) Face detection results of visual measurement, (b) audio-visual feature fusion heatmap w/o AVCL, (c) audio-visual feature fusion heatmap w/i AVCL, (d) detection results w/o AVCL and (e) detection results w/i AVCL.

cases of visual occlusion, the model leverages the audio measurement method based on stGCF described in Section III-A, using the global sound source map to accurately estimate the positions of multiple speakers, thereby partially compensating for the missing visual information. In scenarios where audio information is missing, such as silence in audio, the model relies

on the lightweight face detector in the visual measurement module to identify and track facial regions of speakers, providing complementary information for audio prediction and cross-modal association. Therefore, even in the case of short-term loss of one modality, AVCLNet can still use the other modality to maintain continuous multispeaker tracking and identity association, demonstrating strong cross-modal robustness and generalisation ability. However, there are still some challenges in real-world applications. First, when there is long-term and large-scale occlusion in the scene, the face detector may fail to provide stable visual cues, thereby affecting the performance of cross-modal feature alignment. Second, when the speaker is out of the camera's view for a long time, the lack of visual guidance significantly reduces the accuracy of identity association. Third, if the face detector produces false positives or false negatives, its output will directly affect the accuracy of modality fusion, leading to incorrect audio-visual associations. These issues indicate directions for further improving the robustness and scalability of the model in the future.

To address these limitations, future work will explore strategies to enhance robustness under occlusion and off-screen conditions, such as integrating temporal audio continuity or visual hallucination techniques. Moreover, we aim to scale the framework to larger and more dynamic scenes with more participants. Real-time optimisation and deployment on resource-constrained edge devices are also important directions. Additionally, extending the model to recognise and track nonspeech acoustic events (e.g., footsteps, object collisions) could further broaden its applicability to general audio-visual perception tasks.

5 | Conclusions

This paper proposes AVCLNet, a multimodal multispeaker tracking network using audio-visual contrastive learning. AVCLNet aims to address the challenges caused by cross-modal feature discrepancies, weak sound source localisation ambiguity and frequent identity switch errors, which lead to difficulties in modelling speaker identity consistency and result in unstable tracking trajectories. By employing audio-visual contrastive learning, heterogeneous modal representations are aligned into a unified feature space, promoting cross-modal feature consistency and alleviating feature bias, thereby enhancing the representation capability of identity consistency. Additionally, the vision-guided weak sound source weighted enhancement method, through cross-modal mapping and temporal dynamic weighting, effectively suppresses noise interference and enhances the detectability of weak sound sources. The dual geometric constraint strategy, by combining 2D and 3D spatial geometric information, enhances the robustness of data association in complex motion scenarios and effectively reduces the occurrence of identity switch errors. Experiments on two benchmark datasets demonstrate that AVCLNet outperforms existing methods in core metrics, particularly in scenarios with overlapping speech and noise interference. Future work will explore more lightweight model designs and expand to dynamic open scenarios, promoting the deployment and generalisation of audio-visual multimodal trackers in real-world applications.

Funding

This work was supported by the National Natural Science Foundation of China (62403345), the Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (2024B1212010006), and the Shanxi Provincial Department of Science and Technology Basic Research Project (202403021212174, 202403021221074).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio Speech and Language Processing* 20, no. 2 (2012): 356–370, https://doi.org/10.1109/tasl.2011.2125954.

2. A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social Lstm: Human Trajectory Prediction in Crowded Spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2016), 961–971.

3. X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 8126–8135.

4. J. S. Yoon, T. Shiratori, S.-I. Yu, and H. S. Park, "Self-Supervised Adaptation of High-Fidelity Face Models for Monocular Performance Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019).

5. Y. Feng, S. Yu, H. Peng, Y.-R. Li, and J. Zhang, "Detect Faces Efficiently: A Survey and Evaluations," *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4 (2022): 1–18, https://doi.org/10.1109/tbiom.2021.3120412.

6. Y. Chen, B. Liu, Z. Zhang, and H.-S. Kim, "An End-To-End Deep Learning Framework for Multiple Audio Source Separation and Localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (2022), 736–740.

7. D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust Sound Source Tracking Using srp-phat and 3d Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2020): 300–311, https://doi.org/10.1109/taslp.2020.3040031.

8. X. Qian, M. M, Z. Pan, J. Wang, and H. Li, "Multi-Target Doa Estimation With an Audio-Visual Fusion Mechanism," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (2021), 4280–4284.

9. E. D'Arca, N. M. Robertson, and J. Hopgood, "Person Tracking via Audio and Video Fusion," in *IET Data Fusion and Target Tracking Conference: Algorithms and Applications*, (2012).

10. X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3d Audio-Visual Speaker Tracking With an Adaptive Particle Filter," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (2017), 2896–2900.

11. Y. Li, H. Liu, and H. Tang, "Multi-Modal Perception Attention Network With Self-Supervised Learning for Audio-Visual Speaker Tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2022), 1456–1463.

12. Y. Li, H. Liu, and B. Yang, "Stnet: Deep Audio-Visual Fusion Network for Robust Speaker Tracking," *IEEE Transactions on Multimedia* (2024): 1–13.

13. X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li, "Audio-Visual Cross-Attention Network for Robotic Speaker Tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023): 550–562, https://doi.org/10.1109/taslp.2022.3226330.

14. B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," in *International Conference on Learning Representations*, (2022).

15. A. Berg, M. O'Connor, Kalle, and M. Oskarsson, "Extending gcc-phat Using Shift Equivariant Neural Networks," in *Proceedings of Interspeech*, (2022), 1791–1795.

16. D. H. Shmuel, J. P. Merkofer, G. Revach, R. J. G. van Sloun, and N. Shlezinger, "Deep Root Music Algorithm for Data-Driven Doa Estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (2023).

17. Y. Li, Z. Zhou, C. Chen, P. Wu, and Z. Zhou, "An Efficient Convolutional Neural Network With Supervised Contrastive Learning for Multi-Target Doa Estimation in Low Snr," *Axioms* 12, no. 9 (2023): 862, https://doi.org/10.3390/axioms12090862.

18. A. Liu, J. Guo, Y. Arnatovich, and Z. Liu, "Lightweight Deep Neural Network With Data Redundancy Removal and Regression for Doa Estimation in Sensor Array," *Remote Sensing* 16, no. 8 (2024): 1423, https://doi.org/10.3390/rs16081423.

19. T. Fischer, T. E. Huang, J. Pang, et al., "Qdtrack: Quasi-Dense Similarity Learning for appearance-Only Multiple Object Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 12 (2023): 15.380–15.393, https://doi.org/10.1109/tpami.2023.3301975.

20. S. Kim, I. Petrunin, and H.-S. Shin, "Afjpda: A Multiclass Multi-Object Tracking With Appearance Feature-Aided Joint Probabilistic Data Association," *Journal of Aerospace Information Systems* 21, no. 4 (2024): 294–304, https://doi.org/10.2514/1.i011301.

21. T. Kropfreiter, F. Meyer, D. F. Crouse, S. Coraluppi, F. Hlawatsch, and P. Willett, "Track Coalescence and Repulsion in Multitarget Tracking: An Analysis of Mht, Jpda, and Belief Propagation Methods," *IEEE Open Journal of Signal Processing* 5 (2024): 1089–1106, https://doi.org/10.1109/ojsp.2024.3451167.

22. V. Kılıç, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-Shift and Sparse Sampling-Based smc-phd Filtering for Audio Informed Visual Speaker Tracking," *IEEE Transactions on Multimedia* 18, no. 12 (2016): 2417–2431, https://doi.org/10.1109/tmm.2016.2599150.

23. A. Brutti and O. Lanz, "A Joint Particle Filter to Track the Position and Head Orientation of People Using Audio Visual Cues," in *European Signal Processing Conference*, (2010), 974–978.

24. H. Zhou, M. Taj, and A. Cavallaro, "Target Detection and Tracking With Heterogeneous Sensors," *IEEE Journal of Selected Topics in Signal Processing* 2, no. 4 (2008): 503–513, https://doi.org/10.1109/jstsp.2008.2001429.

25. Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Exploiting the Complementarity of Audio and Visual Data in Multi-Speaker Tracking," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, (2017), 446–454.

26. F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio–Visual Active Speaker Tracking in Cluttered Indoors Environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38, no. 3 (2008): 799–807, https://doi.org/10.1109/tsmcb.2008.922063.

27. Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational Bayesian Inference for audio-visual Tracking of Multiple Speakers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 5 (2019): 1761–1776, https://doi.org/10.1109/tpami.2019.2953020.

28. R. Brunelli, A. Brutti, P. Chippendale, et al., "A Generative Approach to Audio-Visual Person Tracking," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*, (2006), 55–68.

29. J. Wilson and M. C. Lin, "Avot: Audio-Visual Object Tracking of Multiple Objects for Robotics," in *IEEE International Conference on Robotics and Automation*, (2020), 10045–10051.

30. I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, no. 5 (2017): 1086–1099, https://doi.org/10.1109/tpami.2017.2648793.

31. D. Zotkin, R. Duraiswami, and L. Davis, "Joint Audio-Visual Tracking Using Particle Filters," *EURASIP Journal on Advances in Signal Processing* 2002 (2002): 1–11, https://doi.org/10.1155/s1110865702206058.

32. J. Zhao, P. Wu1, X. Liu, et al., *Audio-Visual multi-speaker Tracking with Improved Gcf and Pmbm Filter* (Interspeech, 2022).

33. U. Kirchmaier, S. Hawe, and K. Diepold, "Dynamical Information Fusion of Heterogeneous Sensors for 3d Tracking Using Particle Swarm Optimization," *Information Fusion* 12, no. 4 (2011): 275–283, https://doi.org/10.1016/j.inffus.2010.06.005.

34. I. D. Gebru, S. Ba, and G. E. andR. P. Horaud, "Audio-Visual Speech-Turn Detection and Tracking," in *International Conference on Latent Variable Analysis and Signal Separation*, (2015), 143–151.

35. M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust Multi-Speaker Tracking via Dictionary Learning and Identity Modeling," *IEEE Transactions on Multimedia* 16, no. 3 (2014): 864–880, https://doi.org/10.1109/tmm.2014.2301977.

36. D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-Visual Speaker Tracking With Importance Particle Filters," in *Proceedings of International Conference on Image Processing*, (2003), 1259–1262.

37. N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple Person and Speaker Activity Tracking With a Particle Filter," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (2004), 881–884.

38. H. Liu, Y. Li, and B. Yang, "3d audio-visual Speaker Tracking With a Two-Layer Particle Filter," in *IEEE International Conference on Image Processing*, (2019), 1955–1959.

39. V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering," *IEEE Transactions on Multimedia* 17, no. 2 (2015): 186–200, https://doi.org/10.1109/tmm.2014.2377515.

40. F. Sanabria-Macias, M. Marron-Romera, and J. Macias-Guarasa, "3d Audiovisual Speaker Tracking With Distributed Sensors Configuration," in *European Signal Processing Conference*, (2021), 256–260.

41. H. Liu, Y. Sun, Y. Li, and B. Yang, "3d Audio-Visual Speaker Tracking With a Novel Particle Filter," in *Proceedings of IEEE International Conference on Pattern Recognition*, (2021), 7343–7348.

42. Y. Liu, V. Kılıç, J. Guan, and W. Wang, "Audio–Visual Particle Flow smc-phd Filtering for Multi-Speaker Tracking," *IEEE Transactions on Multimedia* 22, no. 4 (2020): 934–948, https://doi.org/10.1109/tmm.2019.2937185.

43. J. Zhao, P. Wu, X. Liu, Y. Xu, L. Mihaylova, and S. Godsill, "Audio-Visual Tracking of Multiple Speakers via a Pmbm Filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (2022), 5068–5072.

44. Y. Qian, Z. Chen, and S. Wang, "Audio-Visual Deep Neural Network for Robust Person Verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 1079–1092, https://doi.org/10.1109/taslp.2021.3057230.

45. J. Yu, Y. Cheng, and R. Feng, "Mpn: Multimodal Parallel Network for Audio-Visual Event Localization," in *IEEE International Conference on Multimedia and Expo*, (2021), 1–6.

46. Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-Visual Event Localization in Unconstrained Videos," in *Proceedings of the European Conference on Computer Vision*, (2018), 247–263.

47. D. Salvati, C. Drioli, and G. L. Foresti, "Acoustic Source Localization Using a Geometrically Sampled Grid SRP-PHAT Algorithm With Max-Pooling Operation," *IEEE Signal Processing Letters* 29 (2022): 1828–1832, https://doi.org/10.1109/LSP.2022.3199662.

48. X. Xu, T. Qian, Z. Xiao, N. Zhang, J. Wu, and F. Zhou, "Pgsl: A Probabilistic Graph Diffusion Model for Source Localization," *Expert Systems with Applications* 238 (2024): 122028, https://doi.org/10.1016/j.eswa.2023.122028.

49. P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "Improved Feature Extraction for Crnn-Based Multiple Sound Source Localization," in *European Signal Processing Conference*, (2021), 1–5.

50. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is all You Need," *Advances in Neural Information Processing Systems* (2017): 5998–6008, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

51. C. Schymura, B. Bönninghoff, T. Ochiai, et al., "Pilot: Introducing Transformers for Probabilistic Sound Event Localization," *Interspeech* (2021): 2117–2121, https://doi.org/10.48550/arXiv.2106.03903.

52. L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "Crnn-Based Multiple Doa Estimation Using Acoustic Intensity Features for Ambisonics Recordings," *IEEE Journal of Selected Topics in Signal Processing* 13, no. 1 (2019): 22–33, https://doi.org/10.1109/jstsp.2019.2900164.

53. J. Yang, X. Huang, X. Zhang, Q. Zhang, and P. Mei, "Sound Event Localization and Detection Model Based on Multi-View Attention," *Journal of Signal Processing* 40 (2024): 385–395, https://link.springer.com/article/10.1007/s00034-025-03325-0.

54. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning*, (2021), 8748–8763.

55. C. Jia, Y. Yang, Y. Xia, et al., "Scaling up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in *Proceedings of the International Conference on Machine Learning*, (2021), 4904–4916.

56. H. Bao, W. Wang, L. Dong, et al., "Vlmo: Unified Vision-Language Pre-Training With Mixture-Of-Modality-Experts," in *Advances in Neural Information Processing Systems*, (2022).

57. L. Sun, Z. Lian, B. Liu, and J. Tao, "Hicmae: Hierarchical Contrastive Masked Autoencoder for Self-Supervised Audio-Visual Emotion Recognition," *Information Fusion* 101 (2024), https://doi.org/10.1016/j.inffus.2024.102382.

58. I. Tsiamas, S. Pascual, C. Yeh, and J. Serrà, "Sequential Contrastive Audio-Visual Learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (2025).

59. S. Nakada, T. Nishimura, H. Munakata, M. Kondo, and T. Komatsu, "Deteclap: Enhancing Audio-Visual Representation Learning With Object Information," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (2025).

60. G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16.3: An Audio-Visual Corpus for Speaker Localization and Tracking," in *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, (2004), 182–195.

61. X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-Speaker Tracking From an Audio-Visual Sensing Device," *IEEE Transactions on Multimedia* 21, no. 10 (2019): 2576–2588, https://doi.org/10.1109/tmm.2019.2902489.

62. W. Ao, C. Hui, L. Lihao, et al., "Yolov10: Real-Time End-To-End Object Detection," in *Proceedings of the Conference on Neural Information Processing Systems*, (2024).

63. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255.

64. X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cevallaro, "Audio-Visual Tracking of Concurrent Speakers," *IEEE Transactions on Multimedia* 24 (2022): 942–954, https://doi.org/10.1109/tmm.2021.3061800.