Visually Assisted Self-supervised Audio Speaker Localization and Tracking

Jinzheng Zhao[1], Peipei Wu[1], Shidrokh Goudarzi[1], Xubo Liu[1], Jianyuan Sun[1], Yong Xu[2], Wenwu Wang[1]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2]Tencent AI Lab, Bellevue, WA, USA

***Abstract*—Training a robust tracker of objects (such as vehicles and people) using audio and visual information often needs a large amount of labelled data, which is difficult to obtain as manual annotation is expensive and time-consuming. The natural synchronization of the audio and visual modalities enables the object tracker to be trained in a self-supervised manner. In this work, we propose to localize an audio source (i.e., speaker) using a teacher-student paradigm, where the visual network teaches the audio network by knowledge distillation to localize speakers. The introduction of multi-task learning, by training the audio network to perform source localization and semantic segmentation jointly, further improves the model performance. Experimental results show that the audio localization network can learn from visual information and achieve competitive tracking performance as compared to the baseline methods that are based on the audio-only measurements. The proposed method can provide more reliable measurements for tracking than the traditional sound source localization methods, and the generated audio features aid in visual tracking.**

***Index Terms*—knowledge distillation, audio localization, multi-task learning**

I. Introduction

Localizing multiple speakers simultaneously plays a key role in many civilian applications such as speech recognition [1], human-computer interaction [2], and speaker diarization [3]. Audio and visual signals, as two important modalities, can provide complementary information to improve localization robustness and accuracy [4]. For example, if speakers are occluded or disappear from the camera field of view, they can be localized using audio signals; if the audio information is corrupted by background noise and room reverberation, visual data can be used to locate and detect the speakers. Thus information of multiple modalities can work jointly to improve the localization performance. However, the information of the two modalities is not always available concurrently, and thus a good localization system should perform robustly when one modality is missing.

In this paper, we consider the scenario where the visual modality is missing and the audio signal is used to localize speakers. There are several traditional sound source localization methods such as Global Coherence Field (GCF) [5] and MUltiple SIgnal Classification (MUSIC) [6]. With the deep learning methods thriving, some works [7] [8] tackle this problem by training an audio network. However, these works need large amount of annotated training data, which are hard to obtain. For example, datasets in audio speaker localization such as AV16.3 [9] and AVDIAR [3] have only few annotated sequences. The teacher-student paradigm enables the use of a large-scale unlabeled dataset [10] and avoids the need for manual annotation, which is expensive and time-consuming. This paradigm often requires the teacher network to extract pseudo labels and the student network to match the extracted labels. The teacher networks are often selected as pre-trained models, which offer good performance. Compared to the teacher networks, the student networks used often have more light-weighted architectures.

In recent works, audio was used in semantic segmentation [11], depth perception [12], and acoustic scene classification [10] guided by the teacher modality. Inspired by these works, recently, audio has also been used in speaker detection and localization [13], and vehicle localization [14] [15] following the teacher-student paradigm. Compared to audio modality, visual modality is more informative and has the capability of localizing objects accurately in 2D or 3D spaces [16] using color histogram or pretrained face detector, which can teach the audio modality through knowledge distillation. We train an audio network as the student network to track the speakers, guided by visual network (i.e. the teacher network).

Semantic segmentation aims to predict the class labels for each pixel of an image, such as the methods presented in [17], [18] and [19]. Using knowledge-distillation to teach audio network to perform the semantic segmentation task has been explored recently [11] [12]. In the well-known MOT Challenge, one of the tasks, MOTS, combines the tasks of tracking and segmentation [20]. The setting of multi-task shows that the performance of the sub-tasks can be improved via joint training of multiple similar tasks [12]. We hypothesize that joint training for audio localization and audio semantic segmentation could potentially improve the localization performance as classifying pixels belonging to speakers also need the model to infer the positions of the speakers.

If reliable measurements are obtained, Bayesian-based filters can be used to track the objects, such as Particle filter (PF) [21] which is a sequential Monte Carlo algorithm approximating the state distribution by a number of random weighted particles obtained by sequential importance sampling. To demonstrate the idea, we employ particle filter with measurements generated by audio network for speaker tracking, however, other multi-target tracking algorithms such as probabilistic hypothesis density (PHD) filter [22], Multi-target multi-Bernoulli (MeMBer) filter [22] and Poisson multi-Bernoulli mixture (PMBM) filter [23] [24] can also be used.

In this paper, our contributions are two-fold: (1) We propose a method for speaker localization based on the teacher-student paradigm where a visual network is used to teach audio network, allowing us to make use of unlabeled data e.g. as in AV16.3 dataset [9]; (2) We use multi-task learning to further improve the localization accuracy. Compared to the traditional source localization methods, the proposed method can provide more accurate measurements. This method can also improve the tracking performance when the visual modality is unavailable.

II. Proposed Methods

We introduce the architecture of the teacher-student network in Section II-A and the settings of multi-task learning in Section II-B. The teacher-student network with multi-task learning aims to generate audio measurements. The particle filter employs these measurements to estimate the target states, which is discussed in Section II-C.

A. Teacher-Student Network

The architecture of our model follows the teacher-student paradigm, which is shown in Figure 1. We select Dual Shot Face Detector (DSFD) [25] as the teacher network due to its good performance in detecting human faces. With this detector, we can get coordinates of face bounding boxes at every time frame, i.e. $\boldsymbol{b}_k = [x, y, w, h]^T$, where k is the time instant, $[x, y]$ is the top-left coordinate of the bounding box and $[w, h]$ is the width and height of the bounding box, respectively. Bounding boxes whose confidence scores are above a predefined threshold λ are treated as reliable measurements and are converted to coordinates of mouth position,

$$\boldsymbol{o}_k = \boldsymbol{W} \cdot \boldsymbol{b}_k \tag{1}$$

where $\boldsymbol{o}_k = (x_k, y_k)$ is regarded as the pseudo label for training the audio network at time k, $\boldsymbol{W} = [\boldsymbol{I}, \mathrm{diag}(0.5, 0.75)]$ is the conversion matrix from face bounding boxes to mouth positions, as defined in [16].

The audio feature is taken as the input to the student network. Here, we use Generalized Cross Correlation with Phase Transform (GCC-PHAT) as the audio feature. The input will be passed through seven convolutional layers to reduce the feature dimension and extract high-level features. Each convolutional layer is followed by MaxPool, ReLU, Dropout and BatchNorm. As one audio sequence corresponds to three visual sequences captured from different cameras in the AV16.3 dataset [9], to avoid the network being confused about the correspondence of the audio-visual sequences, we input the information of the 2D position of the camera to the network. The 2D position of the camera is converted from the 3D position through camera calibration information provided in the AV16.3 dataset [9]. In addition, there are two fully-connected layers to increase the feature dimension. Finally, two fully-connected layers are used to infer the mouth position $(\hat{x}_k, \hat{y}_k)$ of the speaker based on the high-level features. Sigmoid function is used to normalize $(\hat{x}_k, \hat{y}_k)$ to avoid a large range of output values. We employ the MSELoss as the localization loss ζ_{loc} to allow the teacher network to supervise the student network,

$$\zeta_{loc} = (x_k - \hat{x}_k)^2 + (y_k - \hat{y}_k)^2 \tag{2}$$

where (x_k, y_k) is normalized by a Sigmoid function. At the inference stage, the trained audio network can generate the coordinates of the speakers with only audio input, which can be served as audio measurements $\boldsymbol{Z}$ for tracking, discussed in Section II-C.

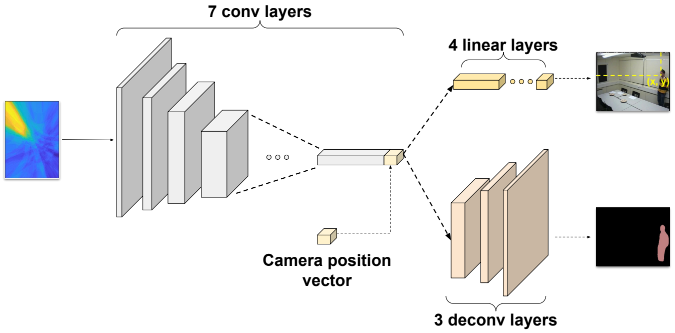

Fig. 1. The architecture of the proposed model. The proposed two tasks have a shared feature extractor. Then the feature goes through linear layers to regress the coordinates and deconvolutional layers to fulfil the semantic segmentation.

B. Multi-Task Learning

We design an auxiliary task to demonstrate that the setting of multi-task learning can improve the localization performance of the student network. In this setting, the student network not only regresses the coordinates of the speaker but classifies the pixels of the speaker and background. We use a pretrained model PSPNet [19] to obtain the pseudo labels $\boldsymbol{l}$ of the semantic segmentation, which has the same size as the input image.

The student network has a separate branch for this auxiliary task. It has a shared group of convolutional layers for the two tasks. After extracting features from the convolutional layers, the network uses three deconvolutional layers to enlarge the dimension of the feature vector to match with the original image size. In the AV16.3 dataset, only moving speakers generate sounds and there are no more other sound sources. Therefore, we modify the PSPNet, by only outputting the class related to the speaker while regarding other classes as the background class. The cross entropy (CE) loss is employed as the segmentation loss ζ_{seg} to supervise the student network:

$$\zeta_{seg} = CE(\hat{\boldsymbol{l}}, \boldsymbol{l}) \tag{3}$$

where $\hat{\boldsymbol{l}}$ is denoted as the predicted labels for each pixel in the input image.

The loss function ζ for multi-task learning is the summation of the localization loss and the segmentation loss:

$$\zeta = \zeta_{loc} + \zeta_{seg} \tag{4}$$

C. Particle Filter

The aim of particle filter is to estimate the state $\boldsymbol{s} = (x, v_x, y, v_y)$ of each speaker with audio measurements, where (x, y) is the 2D coordinates of the speaker's mouth location and (v_x, v_y) is its corresponding velocity. Particle filter uses particles $\boldsymbol{p}_t^{(i)}$ at time step t to represent the state of an object, where i is the particle index. At the start, every particle shares the same weight $w_0^{(i)} = \frac{1}{N}$, where N is the total number of particles. At the prediction stage, the states of the particles are propagated by:

$$\boldsymbol{p}_t^{(i)} = \boldsymbol{F}\boldsymbol{p}_{t-1}^{(i)} + \boldsymbol{q}_t^{(i)} \tag{5}$$

where $\boldsymbol{F}$ is the prediction matrix denoting the velocity-constant dynamic model and $\boldsymbol{q}_t^{(i)}$ is the Gaussian noise with zero mean and covariance $\boldsymbol{Q}$, $\boldsymbol{q}_t^{(i)} \sim \mathcal{N}(0, \boldsymbol{Q})$. In the update stage, the weights of particles are altered by the measurements $\boldsymbol{Z}_t$, which may come from the audio signal, visual signal or signals of other modalities, such as LiDAR information [26] and thermal feature [15],

$$\omega_t^{(i)} \propto g\left(\boldsymbol{Z}_t \mid \boldsymbol{p}_t^{(i)}\right) \tag{6}$$

The measurement likelihood follows Gaussian distribution over the measurement $\boldsymbol{Z}_t$:

$$g\left(\boldsymbol{Z}_t \mid \boldsymbol{p}_t\right) \propto \exp\left[-\left(\boldsymbol{Z}_t - \boldsymbol{p}_t\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Z}_t - \boldsymbol{p}_t\right)\right] \tag{7}$$

where Σ denotes the covariance, indicating the measurement reliability.

The updated state of the speaker is the weighted average over the states of the particles:

$$\boldsymbol{s}_t = \sum_{i=1}^{N} \omega_t^{(i)} \boldsymbol{p}_t^{(i)} \tag{8}$$

The last step is re-sampling, where the particles with large weights are retained and duplicated for the next time step, while the particles with small weights are discarded.

III. Experiments

We introduce the experimental settings and results of audio measurements generation and tracking in this section.

A. Dataset

We use the AV16.3 dataset [9], which is recorded with two 8-microphone arrays with a sampling rate at 16 kHz and three cameras with a sampling rate at 25 fps in an $8.2 \times 3.6 \times 2.4m^3$ meeting room. People in the room are sitting statically, or standing statically, or walking back and forth while speaking at the same time. There are more than 30 sequences in this dataset and only several sequences are annotated with 2D ground truth mouth position of the speakers. To our knowledge, this is the first attempt to use the unlabelled sequences for self-supervised learning on this dataset. We use all sequences of single speaker, containing more than 130,000 frames. When using DSFD [25] to generate pseudo labels for localization, if the frame contains no speaker (i.e. all output confidences are below the pre-defined threshold λ), this frame will be discarded. After processing, we collect 120,417 frames for training and 6,635 for validation. We use sequences 11 and 12, which have the ground truth annotations as the test set, provided in [16].

TABLE I
Measurements obtained from the dataset

	CD_x	CD_y
Teacher network	1.97	2.82
Global Coherence Field (GCF)	17.31	15.07
Mono mel-spectrogram w/o ml	20.47	7.58
8 mel-spectrogram w/o ml	16.84	7.81
GCC-PHAT w/o ml	14.72	7.52
Ours	14.31	7.39

B. Evaluation Metrics

Following [14] and [15], we also use Center Distance (CD) in x and y directions, CD_x and CD_y, to evaluate the localization performance. Center distance denotes the percentage of the localization errors (the distance between the predicted position and the ground truth position of the speaker) in image size.

To evaluate the tracking performance, we use the Optimal Sub-Pattern Assignment (OSPA) [27], defined as

$$E_\rho^{(c)}(\mathbb{M}, \mathbb{N}) = \left(\frac{1}{|\mathbb{N}|}\left(\min_{\pi \in \Pi_{|\mathbb{N}|}} \sum_{i=1}^{|\mathbb{M}|} d^{(c)}\left(m_i, n_{\pi(i)}\right)^\rho + c^\rho\left(|\mathbb{N}| - |\mathbb{M}|\right)\right)\right)^{\frac{1}{\rho}} \tag{9}$$

where $\mathbb{M} = \left\{m_1, m_2, \ldots, m_{|\mathbb{M}|}\right\}$ and $\mathbb{N} = \left\{n_1, n_2, \ldots, n_{|\mathbb{N}|}\right\}$ are two arbitrary finite sets, with $|\cdot|$ being the cardinality of the set, $c > 0$ is the cut-off parameter and $\rho \geq 1$ is the order. $\Pi_{|\mathbb{N}|}$ is the set of permutations on $\{1, 2, \ldots, |\mathbb{N}|\}$. $d^{(c)}\left(m_i, n_{\pi(i)}\right)$ is defined as $\min(c, \|m_i - n_{\pi(i)}\|_2)$. OSPA finds optimal assignment of points in $\mathbb{M}$ and $\mathbb{N}$ and calculates the Euclidean distance of the two matched points. Unmatched points left in $\mathbb{N}$ will result in cardinality error.

C. Implementation Details

To derive the mel-spectrogram, we use the short-time Fourier transform (STFT) with a window size of 1024, a hop size of 256 and 80 frequency bins. For the i-th image frame, we calculate mel-spectrogram of 25 consecutive frames $[i - 12, i + 12]$ in every audio channel, resulting in $80 \times 126 \times 8$ dimensions.

We train our student network using Adam optimizer with 64 batch size and a 5e-5 learning rate. The student network is trained 100 epochs and the early stop mechanism is employed with a patience of 30. The dropout rate is set to 0.1. In extracting localization pseudo labels, the threshold λ is set to 0.5. In semantic segmentation, as the number of pixels of backgrounds is much larger than that of speakers, there exists a class imbalance problem. To mitigate this problem, we set the re-scaling weight of $[1.0, 10.0]$ for the background and the speaker class in the cross entropy loss.

For tracking, the particle filter is initialized with 100 particles and we employ uniform clutter distributed in the image plane with a Poisson rate of 5. For the OSPA, we choose c as 30 and ρ as 2. The detection probability P^{D} is set to 0.98. Every tracker is tested 10 times and the average results are calculated.

D. Comparison Methods for Generating Audio Measurements

Teacher model: The mouth position is estimated by the bounding box generated by DSFD [25] through Equation 1.

Global Coherence Field (GCF): GCC-PHATs of different microphone pairs are aggregated to form the GCF map in x and y dimension, and then the peak value in the GCF map is selected as the estimated location.

Mono mel-spectrogram w/o ml: The mel-spectrogram of the single audio channel is used as the input to the student network without the settings of the multi-task learning.

8 mel-spectrogram w/o ml: The mel-spectrograms of eight audio channels are calculated and stacked together as the input of the student network without the settings of multi-task learning.

GCC-PHAT w/o ml: The GCC-PHAT is used as the input to the student network without the settings of the multi-task learning.

Ours: The GCC-PHAT is used as the input of the student network jointly trained for semantic segmentation.

E. Comparison Methods for Tracking

We use the localization results as measurements for the particle filter to track the speakers. The tracking experiments are conducted on the same sequences as those in Section III-A. We make the following comparisons:

GCF: The **GCF** output is used as measurements (**GCF** in Section III-D).

Ours: The output of our purposed method is used (**Ours** in Section III-D) as measurements.

Visual Only (VO) Tracking: The measurements are estimated by Equation 1 according to the bounding box generated by DSFD [25] (**Teacher** in Section III-D).

Audio-visual (AV) Tracking: The measurements in **Ours** and **VO** are fused. When the information of visual modality is not available (e.g. The face detector fails when people are not facing towards the cameras), the tracker turns to leverage the audio modality.

TABLE II
OSPA Localization Errors

Sequence	Seq 11			Seq 12			Avg
Camera	1	2	3	1	2	3	
GCF	29.63	29.44	28.64	28.52	27.99	27.78	28.67
Ours	23.95	26.89	27.82	26.08	28.41	26.49	26.61
VO	5.72	5.85	4.68	4.27	4.19	3.93	4.77
AV	5.68	5.80	4.63	4.25	4.16	3.91	4.74

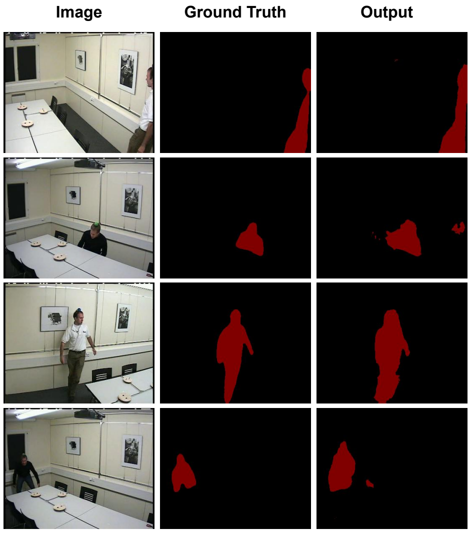

Fig. 2. The output of semantic segmentation by the audio network.

F. Analysis of the Quality of the Audio Measurements

The experimental results for the baseline methods and our proposed methods are listed in Table I. It can be seen that the teacher network has excellent performance and has the capability of teaching the student model to localize speakers. The performance of the student network with mono mel-spectrogram input is worse than that using eight stacked mel-spectrogram, which shows that multi-channel signals provide more positional information and offers more accurate localization results. In addition, the model using GCC-PHAT outperforms the model using mel-spectrogram, indicating that GCC-PHAT is a more powerful feature for localization, as it detects objects based on the mutual information of different microphone pairs. Our proposed method outperforms all the baselines. Compared to **GCC-PHAT w/o ml**, our model shows a lower localization error, indicating the effectiveness of the setting of multi-task learning. As the task of semantic segmentation requires classifying whether the pixel belongs to the speaker, it also needs the model to learn where the speaker is. Thus the semantic segmentation task enhances the performance of the model in speaker localization. As a by-product of the audio network, we visualize the semantic segmentation results in Fig. 2.

Compared to the traditional sound source localization method GCF [16] [28], our proposed data-driven method gives improved performance, showing the advantages of using a large amount of unlabeled training data.

G. Analysis of Tracking results

The tracking results are demonstrated in Table II. Compared to GCF, our method can provide more accurate measurements for tracking, indicating the advantages of leveraging the large amount of unlabelled data for training.

Compared to VO tracking, the tracker using both visual and audio modalities has a lower error. When the highest bounding box confidence is below the pre-defined threshold $\lambda = 0.5$, the tracker turns to use the audio measurements. The measurements generated by our proposed method can help the tracker avoid deviating from the trajectory significantly when the video based face detector has poor detection performance.

IV. Conclusion

In this paper, we have presented a self-supervised learning method for audio speaker localization. The audio network can learn to localize speakers by matching the pseudo labels generated by the teacher model based on video network. The designed auxiliary semantic segmentation task helped to further improve the localization accuracy. The measurements generated by the audio network can benefit visual tracking by serving as a complementary modality. In future works, we will extend our work to the multi-speaker tracking scenarios.

Acknowledgement

This research is sponsored in part by Tencent AI Lab Rhino-Bird Gift Fund and a PhD studentship from the Doctoral College of University of Surrey. The research is also funded in part by the US Army Research Laboratory and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing under the SIGNetS project. It is accomplished under Cooperative Agreement Number W911NF-20-2-0225. The views and conclusions contained in this document are of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the MOD, the U.S. Government or the U.K. Government. The U.S. Government and U.K. Government are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *International Conference on Audio-Visual Speech Processing*, 2003.

[2] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[3] 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*, vol. 40, no. 5, pp. 1086–1099, 2017.

[4] V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2014.

[5] D. C. Marcus, "Acoustic transduction," in *Cell Physiology Source Book*. Elsevier, 2001, pp. 775–791.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[7] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, p. 3418, 2018.

[8] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[9] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[10] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in Neural Information Processing Systems*, vol. 29, pp. 892–900, 2016.

[11] G. Irie, M. Ostrek, H. Wang, H. Kameoka, A. Kimura, T. Kawanishi, and K. Kashino, "Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3961–3964.

[12] A. B. Vasudevan, D. Dai, and L. Van Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in *European Conference on Computer Vision*. Springer, 2020, pp. 638–655.

[13] D. Berghi, A. Hilton, and P. J. B. Jackson, "Visually supervised speaker detection and localization via microphone array," in *IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[14] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7053–7062.

[15] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 612–11 621.

[16] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio–visual sensing device," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[17] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5229–5238.

[18] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[20] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "Mots: Multi-object tracking and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7942–7951.

[21] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[22] R. P. Mahler, *Statistical multisource-multitarget information fusion*. Artech House Norwood, MA, USA, 2007, vol. 685.

[23] Á. F. García-Fernández, J. L. Williams, K. Granström, and L. Svensson, "Poisson multi-bernoulli mixture filter: Direct derivation and implementation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1883–1901, 2018.

[24] J. Zhao, P. Wu, X. Liu, Y. Xu, L. Mihaylova, S. Godsill, and W. Wang, "Audio-visual tracking of multiple speakers via a pmbm filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5068–5072.

[25] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[26] S. Pang and H. Radha, "Multi-object tracking using poisson multi-bernoulli mixture filtering for autonomous vehicles," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7963–7967.

[27] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[28] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Audio-visual tracking of concurrent speakers," *IEEE Transactions on Multimedia*, 2021.