DOI: 10.61782/fa.2025.0262

FORUM ACUSTICUM
EURONOISE 2025

DEEP LEARNING-DRIVEN OBJECT TRACKING FOR ENHANCED ACOUSTIC BEAMFORMING IN DYNAMIC ENVIRONMENTS

Jorge Ortigoso-Narro[1*] **Jose A. Belloch**[2] **Maximo Cobos**[3]

[1] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
[2] Department of Electronic Technology, Universidad Carlos III de Madrid, Spain
[3] Department of Computer Science, Universidad de Valencia, Spain

ABSTRACT

Object tracking and acoustic beamforming are key technologies in applications such as surveillance, human-computer interaction, and robotics. This paper explores integrating deep learning-based object tracking with acoustic beamforming on an embedded device to enhance sound source localization and directional audio capture in dynamic environments. To include depth information, the system was tested with single-camera depth-estimation models and stereo cameras, enabling accurate 3D localization of tracked objects. The system utilizes a planar concentric circular microphone array built with MEMS microphones for compact design and low power consumption, supporting 2D steering capabilities in azimuth and elevation. Positional data from object tracking is processed on the embedded device to dynamically steer the beamforming algorithms, aligning the microphone array's focus with the tracked object's location. The integration of spatial awareness from deep learning trackers with 2D beam steering demonstrates robust performance and adaptability in the presence of moving objects. Experimental evaluations further confirmed that beamforming significantly improves the signal-to-interference ratio, effectively isolating the target source even in dynamic scenarios. This compact design is suitable for teleconferencing, smart home devices, and assistive technologies for the visually impaired, where combining object tracking and beamforming is essential.

Keywords: *Beamforming, Deep Learning, Concentric Circular Arrays, Stereo Vision.*

**Corresponding author*: *jortigos@pa.uc3m.es.*

1. INTRODUCTION

The growing demand for intelligent systems capable of robust sound source localization and directional audio capture has positioned object tracking and acoustic beamforming as pivotal technologies. Applications such as teleconferencing, smart home automation, assistive technologies and other types of human-robot interaction rely on the integration of spatial awareness and audio enhancement to function effectively [1, 2]. Recent advancements in embedded computing and deep learning enable more cohesive solutions, combining visual perception with adaptive audio processing to address these challenges.

The integration of visual tracking and acoustic beamforming has significantly enhanced sound source localization, enabling precise directional audio capture in both classical and modern systems. Devices such as the Microsoft Kinect and Intel RealSense D455 combine RGB cameras, depth sensors, and microphone arrays to spatially focus audio by targeting a speaker's mouth using 3D depth maps to suppress noise. More recent applications, like that by Nagasha et al. [3] which combines face tracking with beamforming for "remote whispering", as well as mobile robots tackling the cocktail party problem [4], utilize deep learning to enhance performance in multi-source environments. Moreover, depth estimation is vital for near-field beamforming accuracy, which can be achieved using either classical structured light meth-

ods (as seen in the Kinect) or modern deep learning approaches such as Monodepth2 [5] and visual-acoustic disparity mapping [6], ensuring adaptability across diverse scenarios.

Real-time audio-visual integration on embedded platforms is achieved through hardware-software co-design, balancing computational efficiency with performance. Depth maps derived from visual sensors provide spatial coordinates that dynamically steer beamforming algorithms, enabling precise audio focus in compact, low-power devices. Studies such as [7] and [8] demonstrate how model pruning, quantization, and hardware acceleration (e.g., GPU/TPU offloading) overcome resource constraints. These optimizations empower applications in smart homes, teleconferencing, and assistive technologies, where latency-critical noise suppression and spatial adaptability are required.

Complementing these hardware-software optimizations, MEMS microphone arrays serve as the acoustic backbone of embedded audio-visual systems, enabling the low-power, high-density configurations required for real-time beamforming [9]. Their compact form factor and minimal energy consumption align seamlessly with the resource constraints of embedded platforms discussed earlier, while their ability to be densely packed into arrays ensures precise directional audio capture. This work focuses on a system that pairs deep learning-based visual tracking with acoustic beamforming. The visual component identifies and localizes objects in 3D space using stereo cameras or monocular depth estimation, while simultaneously classifying targets or other noise sources. This spatial and semantic data is then used to steer a compact MEMS microphone array, which performs beamforming to capture directional audio aligned with the tracked object's position. By dynamically updating the beamformer's focus based on real-time visual inputs, the system maintains accurate audio capture even as targets move.

The main contribution of this work is the development of a compact and efficient embedded platform that integrates deep learning-based 3D object tracking with dynamic MEMS beamforming, enabling robust and adaptive directional audio capture in real-world dynamic environments.

2. SYSTEM DESIGN

At a high level, the architecture comprises three core components: (1) a visual perception module using a stereo or monocular camera, (2) a planar concentric circular MEMS microphone array for audio capture, and (3) a NVIDIA Jetson Orin Nano embedded processor responsible for coordinating tracking, classification, and beam steering. The components are interconnected via a low-latency pipeline, enabling closed-loop interaction between vision and acoustics. Fig. 1 represents the high-level block diagram design of the interconnected systems.

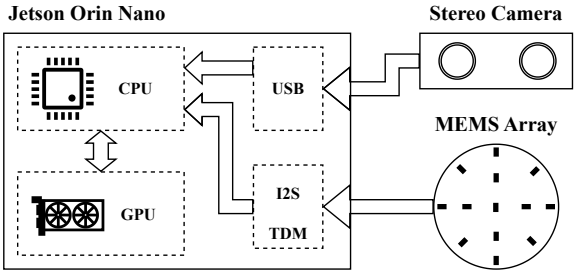

Figure 1. High level block diagram design.

2.1 Vision system design

In order to provide visual-aid information to the system, low-cost USB cameras were chosen for the design. These cameras not only offer a budget-friendly solution but also deliver sufficient resolution and frame rates for real-time object tracking and depth estimation.

To achieve accurate tracking and depth measurement, we performed tests with two different approaches. First, a stereo camera is employed to compute disparity maps, which are converted into 3D spatial coordinates. Second, a single-camera depth-estimation model based on pre-trained neural networks is employed. This model processes the 2D images captured by the monocular camera and infers depth information from visual cues, such as texture gradients and object occlusions. The latter approach allows for a more compact final system since it only requires one camera, but relies on having enough training data of the deployment environments as well as access to hardware acceleration.

The camera module is integrated with the embedded device through a dedicated image processing pipeline where video streams are pre-processed to correct for lens distortions and to normalize illumination, ensuring consistent input quality. These pre-processed images are subsequently fed into the deep learning models that detect and track objects, providing dynamic positional data. The object tracking system continuously updates the position of

targets, which in turn guides the beamforming algorithms. This dynamic integration allows the system to adjust the focus of the microphone array in real time, aligning with the tracked object's location for improved audio capture in cluttered and noisy environments.

2.2 Hardware and array design

A concentric circular microphone array design was chosen due to its symmetry, allowing flexible steering capabilities and good frequency response. This configuration is particularly advantageous for applications requiring omnidirectional sound capture and beamforming, as the circular arrangement ensures uniform spatial resolution in all directions. Additionally, the concentric structure enables the array to operate effectively across a wide range of frequencies, making it suitable for tasks such as sound source localization, speech enhancement, and acoustic scene analysis. The designed array, shown in Fig. 2.2, is comprised of two concentric circular rings of radius $2.5\,\mathrm{cm}$ and $4.5\,\mathrm{cm}$ respectively, as well as a central microphone.

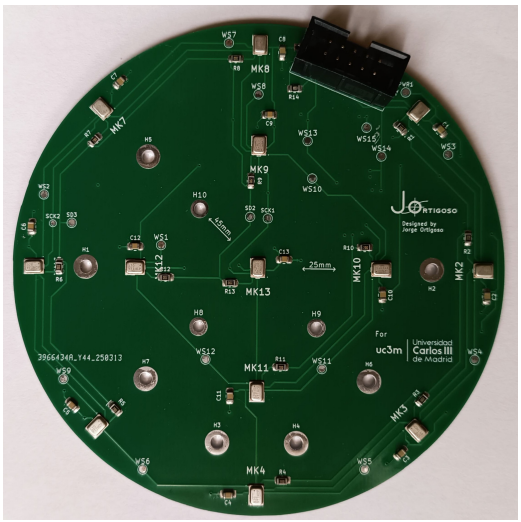

Figure 2. Final designed and assembled printed circuit board.The array design contains $R = 3$ rings of radii $\rho = \{0, 2.5, 4.5\}\,\mathrm{cm}$. The microphones are equally spaced along each ring: for $r = 2$, $\Delta\phi = \pi/2\,\mathrm{rad}$, and for $r = 3$, $\Delta\phi = \pi/4\,\mathrm{rad}$.

To construct the array, ICS52000 MEMS microphones were utilized. These microphones are designed to operate in a daisy-chain configuration, which simplifies data acquisition by consolidating the digitized audio data from all microphones into a single serial stream. Specifically, a modified I2S interface is employed, where the data from each microphone is multiplexed in the time domain using a simple TDM synchronization pulse, ensuring efficient and synchronized transmission.

The PCB designed to house the microphones was meticulously engineered to maintain signal integrity. Termination resistors were placed adjacent to the microphone data line pins, and controlled impedance tracks were traced to match the selected clock frequency. To minimize signal propagation delays, the clock and data lines were arranged in a branch-style topology. Furthermore, signal buffers were incorporated to ensure sufficient driving strength for both the host interface and the microphones, preserving signal amplitude and integrity. High-bandwidth, rail-to-rail, unity-gain stable OPA2810 operational amplifiers were specifically chosen for this purpose, providing robust performance and reliability in the array's operation [10].

2.3 Target tracking

To calculate the steering vector and beamforming coefficients for the microphone array, the elevation (θ_0) and azimuth (ϕ_0) angles are calculated from the estimated target position in space following Equations 1 and 2. The geometric model of the system is depicted in Fig. 3

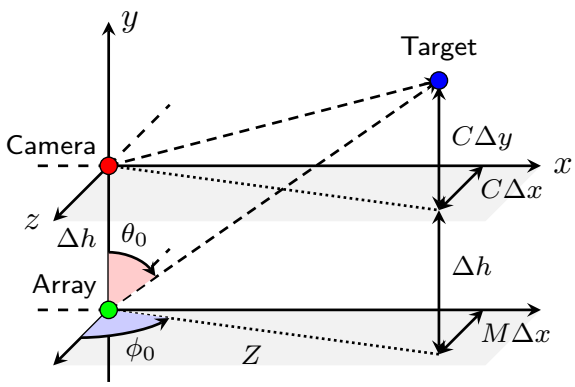

Figure 3. System geometry.

$$\phi_0 = \pi - \arctan\left(\frac{M\Delta x}{Z}\right)\,[\mathrm{rad}] \tag{1}$$

$$\theta_0 = \arctan\left(\frac{C\Delta y + \Delta h}{Z}\right)\,[\mathrm{rad}] \tag{2}$$

The target tracking subsystem enables real-time 3D localization of objects to dynamically guide the acoustic beamforming process. To achieve robust detection and spatial estimation, we utilize a fine-tuned YOLOv10n network for detecting common audio sources like persons and speakers, to estimate bounding boxes of target objects in sequential video frames. By focusing on the centroids of the detected bounding boxes, the system extracts 2D positional data, which is fused with depth information from either monocular depth estimation models or stereo camera disparity maps to resolve 3D coordinates. The YOLOv10n architecture was selected for its computational efficiency, leveraging nested sparse gradients and a decoupled head design to reduce latency while maintaining detection accuracy in cluttered environments [11].

3. EXPERIMENTS AND RESULTS

To evaluate the complete system, several experiments were performed. First, different depth-estimation approaches were assessed in terms of accuracy and latency. Next, the full beamforming pipeline was tested in a controlled environment inside an anechoic chamber. A sampling frequency of 8 kHz was selected to reduce latency by limiting the number of samples.

3.1 Depth estimation

Classic stereoscopic cameras determine depth by using binocular disparity and triangulation. They work very well in environments with plenty of texture but require accurate calibration and can struggle in areas lacking detail [12]. On the other hand, advanced infrared (IR) systems use active illumination to boost performance in low-texture or low-light conditions, although this improvement comes with increased costs due to the need for IR emitters, specialized sensors, and additional synchronization hardware. Monocular deep learning methods sidestep much of this hardware complexity by inferring depth from single images through learned data priors. However, these methods often produce results that lack absolute scale, depend heavily on the quality of training data, and demand significant computational resources. In controlled, metric-critical applications such as robotics, stereoscopic IR systems are preferred, while monocular approaches offer a more cost-effective and scalable option for consumer devices, even if they sometimes sacrifice accuracy and generalization [13].

We benchmark some of the most popular methods with our system to evaluate their depth estimation performance and latency. Table. 1 presents the average timings of 30 runs for each method, along with their standard deviations, measured in seconds.

Table 1. Average timings for the tested depth estimation methods expressed as mean±standard deviation.

Method	Timing (seconds)
CREStereo [14]	1.71 ± 0.45
Depth-AnythingV2 [15]	1.33 ± 0.20
Depth-AnythingV2 (metric)	0.15 ± 0.09
Depth-pro [16]	32.32 ± 11.44
RAFT-Stereo [17]	5.67 ± 1.07
RT-Mono-Depth [18]	0.16 ± 0.08
RT-Mono-Depth (small)	0.11 ± 0.09
SGBM [19] (filtered)	1.73 ± 0.26
StereoNet [20]	1.26 ± 0.39

As shown in the comparison, several deep learning-based models compute depth maps significantly faster than the standard SGBM algorithm. This difference in speed is mainly due to the additional processing steps required by SGBM, including residual filtering and weighted least squares, which are necessary when using a low-cost visual spectrum stereo camera. Figure 4 shows the resulting depth maps, from left to right: filtered SGBM, the fastest deep learning model (RT-Mono-Depth-Pro), and the slowest method overall (Depth-Pro). Although the evaluated environment is particularly challenging (due to the likely very limited presence of anechoic chambers in the training data) the larger deep learning-based models are still able to capture fine details in the scene, but struggle with real distances. Nonetheless, we chose to use filtered SGBM as the depth estimation method for the following experiments because of its robustness.

3.2 Beamforming performance

To evaluate the integrated system in isolation, its performance was first assessed under controlled conditions. In an anechoic chamber, a dual-source setup was arranged with two household loudspeakers serving as sound sources and a dummy head simulating the target source (illustrated in Fig. 5), ensuring a noise-isolated, reflection-free environment. To simulate dynamic conditions, the

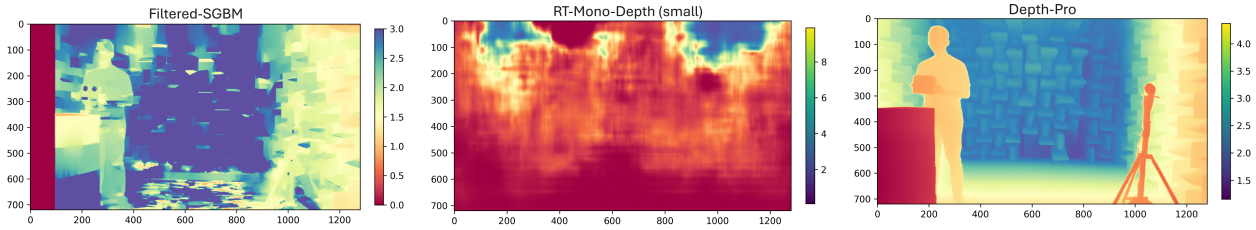

Figure 4. Depth maps obtained for Filtered-SGBM, RT-Mono-Depth and Depth-pro methods. The colorbars show the estimated depth in meters.

experiment was repeated with one moving source, where a person carrying one of the loudspeakers walked at a normal pace toward the fixed source.

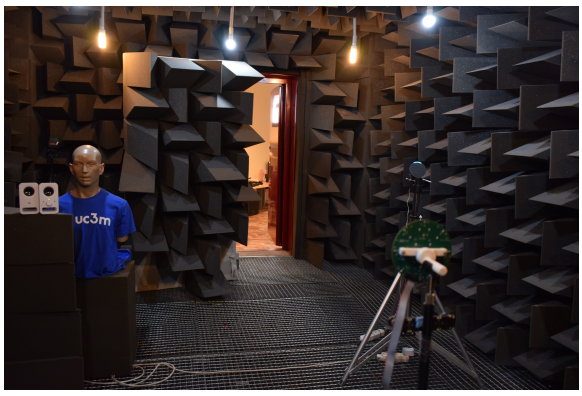

Figure 5. Experimental setup for the static anechoic test.

To extract relevant data, we conducted experiments using two pairs of acoustic signals. In the first scenario, the sources emitted pure tones at 2 kHz and 3 kHz, respectively. In the second one, the 3 kHz tone was replaced by broadband noise, specifically a voice sequence from the TIMIT dataset [21]. Object detection was employed to localize the target of interest (i.e., person-like objects), and beamforming was performed using a time-domain delay-and-sum algorithm. The steering vector was computed based on the estimated direction of arrival (DoA). Figure 6 shows the visual and orientation information extracted from an arbitrary video frame captured during the source emission.

In the two-tone experiment, the signal-to-interference ratio (SIR) between the 2 kHz and 3 kHz signals was measured over time to evaluate the system's ability to separate sources. Fig. 7 (top) displays ΔSIR results for static and

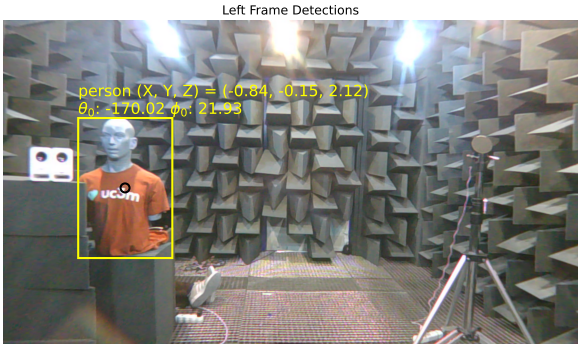

Figure 6. Target detection employed for the static experiment.

dynamic conditions, while the bottom plot shows the estimated azimuth angle of the moving source. Due to variability in source emission stability, the analysis compares differences in SIR between non-beamformed and beamformed signals rather than absolute values. The beamformed results indicate an improvement in SIR, reflecting the system's capacity to estimate the target source position and partially isolate the desired signal. In the dynamic case it can clearly be seen how when the moving target approaches the fixed noise source, the SIR declines as the spatial proximity of the two emitters causes the beam to align with both simultaneously.

For the experiment where the source of interest emitted speech audio, the SIR metric was calculated by by following Eq. 3.

$$\Delta\mathrm{SIR} = \mathrm{SIR}_{\text{Beamformed}} - \mathrm{SIR}_{\text{Not-beamformed}}\ [\mathrm{dB}]\,, \quad (3)$$

$$\mathrm{SIR} = 10\log_{10}\left(\frac{\mathrm{PWR}_{\text{speech}} - \mathrm{PWR}_{\text{int}}}{\mathrm{PWR}_{\text{int}}}\right)[\mathrm{dB}]$$

where ΔSIR quantifies the improvement in the SIR achieved through beamforming. The SIR is computed

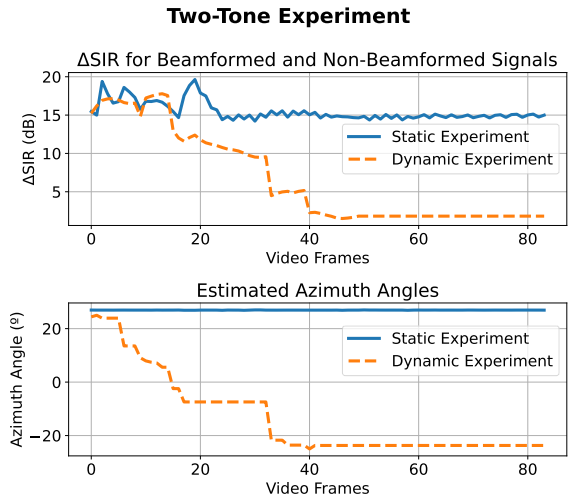

Figure 7. Top plot shows the differences between the beamformed and non-beamformed SIR for the dynamic and static experiments for the two tone case. Bottom plot shows the estimated azimuth angle for the target source.

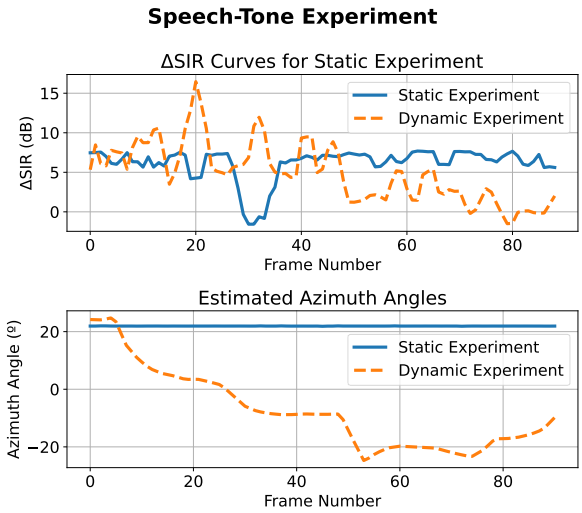

Figure 8. Top plot shows the differences between the beamformed and non-beamformed SIR for the dynamic and static experiments where the signal emmitted by the target of interest was speech. Bottom plot shows the estimated azimuth angle for the target source.

based on the difference between the estimated power of the speech signal and that of the interference, normalized by the interference power. The interference power $\mathrm{PWR}_{\mathrm{int}}$ is estimated by integrating the power spectral density within a 100 Hz bandwidth centered at the known interference tone frequency band.

4. CONCLUSIONS

In conclusion, this work demonstrates the effectiveness of combining visual depth estimation with beamforming techniques for audio source separation. Our experimental results in an anechoic chamber show that the system successfully improves signal-to-interference ratio in both static and dynamic scenarios, with particularly notable performance when spatial separation between sources is maintained. While deep learning-based depth estimation methods showed promising speed advantages, filtered SGBM provided the most robust performance despite its higher computational requirements, especially in challenging visual environments that were likely underrepresented in training datasets. These findings highlight the importance of selecting appropriate depth estimation techniques based on application-specific requirements for reliability versus processing speed.

5. ACKNOWLEDGMENTS

The authors express their sincere gratitude to Ricardo Moreno and Jesus Peña Rodríguez for their invaluable support and assistance, and to Luis A. Azpicueta Ruiz for providing access to the anechoic chamber at the Signal Theory and Communications Department of the University Carlos III de Madrid. Their help and encouragement were instrumental in completing this study. This work has been supported by Grants TED2021-131003B-C21 and TED2021-131401A-C22 funded by MCIN/AEI/10.13039/501100011033 and by the "EU Union NextGenerationEU/PRTR", as well as by Grants PID2022-137048OB-C41 and PID2022-137048OA-C43 funded by MICIU/AEI/10.13039/501100011033 and "ERDF A way of making Europe"

6. REFERENCES

[1] S.-C. Hsia, S.-H. Wang, C.-M. Wei, and C.-Y. Chang, "Intelligent object tracking with an automatic image

zoom algorithm for a camera sensing surveillance system," *Sensors*, vol. 22, no. 22, 2022.

[2] T. Price, D. Howard, A. Lewis, and A. Tyrrell, "Adaptive microphone array beamforming for teleconferencing using vhdl and parallel architectures," in *Proceedings of the Seventh Euromicro Workshop on Parallel and Distributed Processing. PDP'99*, pp. 13–18, 1999.

[3] H. Mizoguchi, Y. Tamai, K. Shinoda, S. Kagami, and K. Nagasghima, "Visually steerable sound beam forming system based on face tracking and speaker array," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 977–980 Vol.3, 2004.

[4] Z. Shi, L. Zhang, and D. Wang, "Audio–visual sound source localization and tracking based on mobile robot for the cocktail party problem," *Applied Sciences*, vol. 13, no. 10, 2023.

[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," October 2019.

[6] W. Sun and L. Qiu, "Visual-assisted sound source depth estimation in the wild," 2022.

[7] P.-L. Asselin, V. Coulombe, W. Guimont-Martin, and W. Larrivée-Hardy, "Replication study and benchmarking of real-time object detection models," 2024.

[8] J. Chen, J. Chen, H. Min, and X. Wang, "Real-time embedded implementation of adaptive beamforming for medical ultrasound imaging," in *2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, pp. 356–360, 2016.

[9] J. Ortigoso Narro, R. Moreno, D. de la Prida Caballero, M. Raiola, and L. Azpicueta-Ruiz, "64-microphone module for a massive acoustic camera," 09 2024.

[10] Texas Instruments, "OPA2810 Data Sheet." `https://www.ti.com/product/OPA2810`, 2023. [Accessed 2025-03-06].

[11] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[12] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second ed., 2004.

[13] Z. Zhang, Y. Zhang, Y. Li, and L. Wu, "Review of monocular depth estimation methods," *Journal of Electronic Imaging*, vol. 34, Mar. 2025.

[14] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16263–16272, 2022.

[15] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[16] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv*, 2024.

[17] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *International Conference on 3D Vision (3DV)*, 2021.

[18] C. Feng, Z. Chen, C. Zhang, W. Hu, B. Li, and L. Ge, "Real-time monocular depth estimation on embedded systems," in *IEEE International Conference on Image Processing, ICIP 2024, Abu Dhabi, United Arab Emirates, October 27-30, 2024*, pp. 3464–3470, IEEE, 2024.

[19] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[20] X.-S. Contributors, "X-StereoLab stereo matching and stereo 3d object detection toolbox." `https://github.com/meteorshowers/X-StereoLab`, 2021.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.