World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3089

MONOCULAR 3D BINAURAL LOCALIZATION AND DYNAMIC TRACKING FOR EAR-SIDE ACTIVE NOISE CONTROL IN VEHICLE CABINS

Cong Zhang
College of Automotive and Energy Engineering, Tongji University, Shanghai 201804, China.

Abstract: Accurate three-dimensional ear-position tracking is a prerequisite for ear-side active noise control in vehicle cabins, because natural head motion can directly shift the control target region and degrade the spatial consistency between the actual ears and the modeled control points. To address this problem, this study proposes a monocular-vision-based method for binaural three-dimensional localization and dynamic tracking in cabin scenes. A unified mapping among the pixel coordinate system, camera coordinate system, and cabin coordinate system is first established through camera calibration and geometric modeling. Facial landmark detection is then used to infer the two-dimensional locations of the left and right ears, after which cabin-feature-constrained monocular depth estimation is introduced to recover ear-region depth in a spatially aligned manner. The binaural three-dimensional coordinates are further refined through head-pose compensation, temporal filtering, and short-term prediction, so that the final output can be directly used as a control-oriented ear-state sequence. Multi-condition experiments under different illumination, occlusion, and head-pose variations show that the proposed method maintains good localization accuracy and trajectory continuity in all six tested cases. The root-mean-square error of binaural three-dimensional localization ranges from 18.40 to 25.57 mm, while the mean interaural distance error remains within 0.37 to 3.33 mm. Even under adverse conditions such as weak illumination and partial occlusion, the processed ear-state output remains continuously available to the downstream active noise control interface. These results indicate that the proposed method provides a low-cost and practically deployable solution for monocular binaural ear tracking in vehicle cabins and can serve as an effective perception front-end for ear-side active noise control.
Keywords: Monocular vision; Binaural localization; Ear tracking; Cabin perception; Active noise control; Intelligent cockpit

1 INTRODUCTION

In-vehicle acoustic comfort is increasingly shifting from global sound-field reduction to more localized and listener-centered regulation. In ear-side active noise control applications, the spatial positions of the occupant's ears are directly related to the control target region, the effective error-point location, and the geometric relationship between the ears and nearby acoustic actuators. Once the occupant turns the head, changes posture, or leans laterally during driving or riding, a fixed-ear assumption is no longer valid, and the mismatch between the actual ear positions and the modeled control points may lead to target drift and performance degradation. This issue becomes particularly critical in local ear-side control scenarios because the useful quiet zone is spatially compact and highly sensitive to geometric variation[1–3].

For this reason, accurate and continuously available ear-position information has become a key prerequisite for practical ear-side active noise control in vehicle cabins. However, direct ear tracking in monocular cabin images remains challenging. Compared with larger facial regions, the ear occupies a smaller image area, exhibits weaker local texture, and is more easily affected by hairstyle, local occlusion, illumination fluctuation, and viewpoint change. In addition, the camera installation in a real cabin is constrained by the cockpit layout, which further limits observation flexibility and makes stable three-dimensional ear localization more difficult.

To address these challenges, this paper proposes a monocular-vision-based method for binaural three-dimensional localization and dynamic tracking in vehicle cabins. The proposed framework establishes a unified mapping from image-plane observations to cabin-coordinate binaural positions through camera calibration and geometric modeling, infers binaural two-dimensional ear locations from facial landmarks, recovers ear-region depth through cabin-feature-constrained monocular depth estimation, and further improves output continuity by introducing head-pose compensation, temporal filtering, and short-term prediction. In this way, the final output is transformed from frame-wise visual observations into a control-oriented ear-state sequence suitable for downstream ear-side active noise control.

The remainder of this paper is organized as follows. Section 2 reviews related work on in-cabin visual perception, head-pose estimation, three-dimensional cabin sensing, and ear-position-aware active control. Section 3 introduces the proposed methodology. Section 4 presents the experimental setup and evaluation metrics. Section 5 reports the experimental results and discussion. Finally, Section 6 concludes the paper.

2 LITERATURE REVIEW

Research on in-cabin visual perception has evolved from early driver monitoring tasks toward more comprehensive occupant spatial-state estimation. Earlier studies mainly focused on vigilance monitoring, driver inattention analysis, and gaze-related behavior understanding, showing that visual sensing has become a fundamental part of intelligent cockpit perception systems[4–9]. These studies established the practical relevance of cabin-facing visual sensing and demonstrated that in-vehicle monitoring could be implemented under realistic driving conditions.

As this field developed, the technical focus gradually moved from coarse face detection to continuous head-pose estimation and three-dimensional state recovery. Representative studies have examined robust head-pose estimation for driver assistance, survey-level methodological categorization of head-pose estimation, monocular head-pose modeling, descriptor-fusion-based driver pose estimation, and reduced-feature-set driver head-pose prediction[10–14]. Collectively, these studies indicate that in-cabin perception has already established a relatively mature basis for extracting facial geometry and spatial pose information under realistic vehicle conditions.

At the same time, three-dimensional in-cabin perception methods have continued to expand through RGB-D sensing, low-cost depth cameras, and full-body or occupant-shape estimation. Continuous gaze-zone estimation with RGB-D cameras, multi-frame point-cloud-based head-pose and gaze-zone estimation, and low-cost depth-camera occupant head tracking have all shown that geometric perception in vehicle cabins can be achieved with increasing robustness[15–17]. More recent work has further extended the target from the head alone to full in-vehicle human pose and body-shape estimation, while monocular-depth-related research has also begun to enter the in-cabin domain through dedicated benchmarks[18–20]. These developments improve the feasibility of low-cost three-dimensional cabin perception, but they do not directly solve the problem of continuous binaural ear tracking for active control.

From the perspective of acoustic applications, studies in active noise control and active headrest systems have shown that listener-centered local control is highly sensitive to the geometric relationship between the ears and the secondary sources, especially under head motion[21–23]. This means that the effectiveness of ear-side active control depends not only on the control algorithm itself but also on whether the ear positions can be estimated reliably and updated continuously during natural head movement.

More specifically, recent studies have examined the accuracy requirements of ear positioning for active control of road noise in a car and have further demonstrated the integration of a depth-camera-based ear-positioning system with an active headrest platform[24-25]. These studies confirm the practical importance of ear-state perception, but they still leave a key engineering problem insufficiently addressed: how to generate cabin-referenced, temporally continuous, and control-usable binaural three-dimensional ear trajectories from a low-cost monocular visual system under realistic illumination variation, local occlusion, and head-pose change.

Against this background, the present study focuses on monocular-vision-based binaural three-dimensional localization and dynamic tracking for vehicle cabins. Compared with depth-camera-based solutions, the proposed method aims at a lower-cost and more easily deployable perception front-end, while still maintaining the geometric consistency and temporal continuity required by downstream ear-side active noise control.

3 METHODOLOGY

3.1 Overall Framework and Control-Oriented Output Definition

The objective of the proposed method is not merely to detect ear regions in monocular cabin images, but to generate temporally continuous and geometrically consistent left-ear and right-ear trajectories in the cabin coordinate system for downstream ear-side active noise control. Accordingly, the overall pipeline is organized into five stages: camera calibration and geometric mapping, binaural two-dimensional inference from facial landmarks, cabin-feature-constrained monocular depth estimation, binaural three-dimensional coordinate reconstruction, and dynamic temporal robustification.

For frame k, the visual observation output is first expressed as

$$\mathbf{z}_v(k)=\left[t_k,\quad \mathbf{p}_{w,L}^{\mathrm{T}}(t_k),\quad \mathbf{p}_{w,R}^{\mathrm{T}}(t_k),\quad \gamma_L(k),\quad \gamma_R(k)\right]^{\mathrm{T}}, \tag{1}$$

where t_k denotes the visual timestamp, $\mathbf{p}_{w,L}^{\mathrm{T}}(t_k)$ and $\mathbf{p}_{w,R}^{\mathrm{T}}(t_k)$ are the reconstructed left-ear and right-ear positions in the cabin coordinate system, and $\gamma_L(k)$ and $\gamma_R(k)$ represent the validity indicators of the two ear observations. Since the downstream acoustic control module does not directly consume frame-wise raw observations, the final control-oriented state is further defined as

$$\mathbf{z}_c(n)=\begin{bmatrix}\hat{\mathbf{p}}_{w,L}(t_n)\\ \hat{\mathbf{p}}_{w,R}(t_n)\end{bmatrix}, \tag{2}$$

where $\hat{\mathbf{p}}_{w,L}(t_n)$ and $\hat{\mathbf{p}}_{w,R}(t_n)$ are temporally aligned, filtered, and short-term predicted coordinates at the control instant t_n. In this way, the visual front-end provides not isolated detections, but a control-usable binaural state sequence with unified spatial reference and temporal continuity.

3.2 Camera Calibration and Geometric Mapping

To convert image-plane observations into physical three-dimensional ear coordinates, a unified mapping among the pixel, camera, and cabin coordinate systems is established. The camera is modeled by a pinhole projection with radial distortion correction. After undistortion, the homogeneous pixel vector is written as

$$\mathbf{p}_d=[u_d,v_d,1]^{\mathrm{T}}. \tag{3}$$

Let the ear point in the camera coordinate system be denoted by

$$\mathbf{p}_c=[X_c,Y_c,Z_c]^{\mathrm{T}}, \tag{4}$$

then the back-projection relationship is

$$\mathbf{p}_c=Z_c\mathbf{K}^{-1}\mathbf{p}_d, \tag{5}$$

where $\mathbf{K}$ is the intrinsic matrix of the calibrated monocular camera. The corresponding cabin-coordinate position is obtained by rigid transformation:

$$\mathbf{p}_w=\mathbf{R}_{wc}\mathbf{p}_c+\mathbf{t}_{wc}, \tag{6}$$

where $\mathbf{R}_{wc}$ and $\mathbf{t}_{wc}$ denote the rotation matrix and translation vector from the camera frame to the cabin frame.

The monocular camera used in this study is a Logitech C920Pro. Calibration is performed at a resolution of 640×480 using a planar checkerboard and Zhang's method [26]. The final intrinsic parameters adopted in the experiments are f_x=592.806 px, f_y=591.961 px, u_0=334.830 px, and v_0=261.092 px. The radial distortion parameters are $k_1=8.070\times10^{-4}$ and $k_2=-0.0984$. The overall mean reprojection error is 0.1535 px, indicating that the calibrated monocular geometric model is sufficiently accurate for subsequent binaural three-dimensional reconstruction (Figure 1).

(a) Camera calibration scene (b) Binaural tracking scene

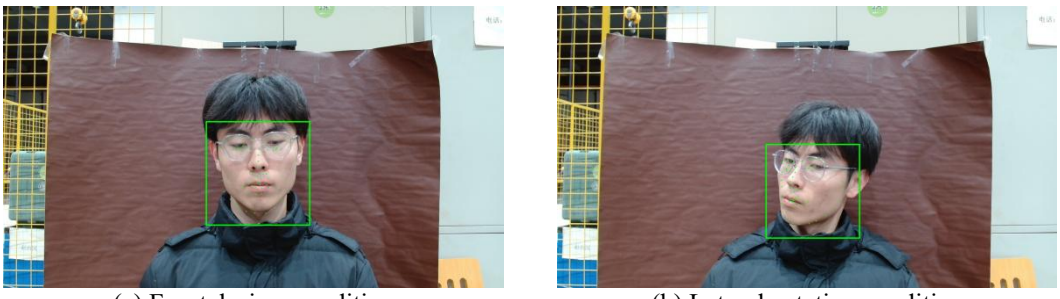
Figure 1 Monocular Visual Platform and Representative Calibration and Binaural-Tracking Scenes

3.3 Binaural Two-Dimensional Inference Based on Facial Landmarks

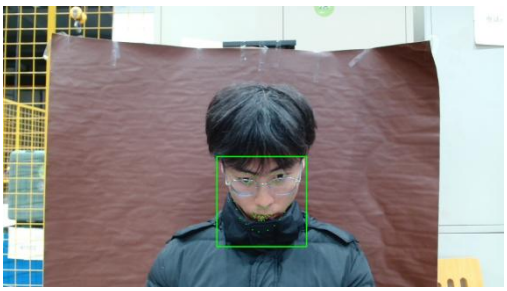
Direct ear detection in cabin images is susceptible to weak ear texture, hairstyle interference, partial occlusion, and large head rotation. To improve robustness, binaural two-dimensional inference is performed using facial geometric structure instead of relying solely on raw ear appearance. The method first detects the face and extracts stable facial landmarks using a landmark-based facial analysis front-end[27-29]. These landmarks provide structural constraints from the eyes, nose, mouth, and facial contour, which are then used to infer the lateral regions corresponding to the left and right ears.

Let the inferred binaural image-plane positions be denoted by (u_L,v_L) and (u_R,v_R). Around these locations, left-ear and right-ear regions of interest, R_L and R_R, are constructed to support subsequent depth sampling. Compared with single-point ear detection, the ROI-based representation is more robust to local landmark jitter and minor contour disturbances, while still preserving the directional and structural information required for binaural localization.

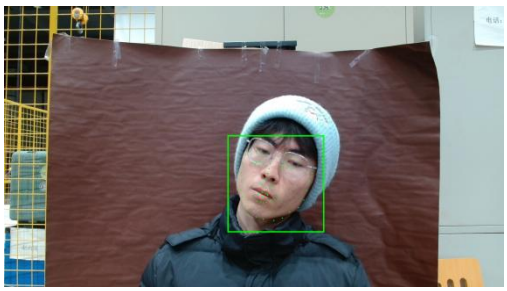
In practical cabin scenes, the inferred ROIs are also subjected to image-boundary truncation and basic validity screening. Therefore, the output of this stage is not just a pair of image coordinates, but a binaural two-dimensional localization result with region-level support, which serves as the spatial anchor for later depth recovery (Figure 2).



(a) Frontal-view condition (b) Lateral-rotation condition



(c) Pitch-variation condition



(d) Appearance-disturbance and pose-coupling condition

Figure 2 Representative Face Detection and Facial Landmark Extraction Results under Different Cabin Conditions

3.4 Cabin-Feature-Constrained Monocular Depth Estimation

The proposed method does not use monocular depth estimation as an isolated vision task. Instead, the depth field is introduced as an intermediate geometric representation for stable ear-region depth recovery. To this end, a cabin-feature -constrained monocular depth module is constructed. It combines image-aligned dense depth estimation with cabin-background geometric cues, so that the resulting depth map is more suitable for later ear-state reconstruction than directly applying a generic depth predictor.

Let the input RGB image be I, and let the depth-estimation network output the aligned depth map $D(u,v)$. Following recent progress in fast monocular depth estimation [30], the depth module is organized in a multi-scale manner so that both local detail and broader cabin geometry can be retained. In addition, static cabin background structures are used as supplementary constraints to improve geometric consistency in the target field of view.

A key point is that the final ear depth is not taken from a single pixel. Instead, a local ear-region depth sampling strategy is used. For each ear ROI, candidate depth values are gathered within the valid region and then robustly aggregated to obtain $Z_{c,L}$ and $Z_{c,R}$. This strategy reduces the influence of isolated noisy pixels, weak-texture artifacts, and local depth discontinuities near the facial contour. It therefore improves the temporal stability of ear-depth estimation, especially under weak illumination and partial occlusion.

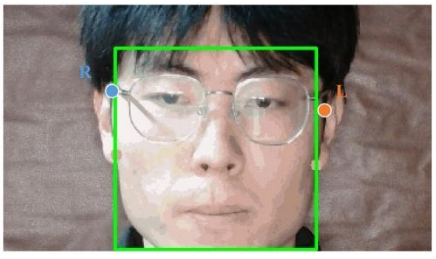
3.5 Binaural Three-Dimensional Reconstruction and Dynamic Tracking

3.5.1 Joint three-dimensional reconstruction

With the inferred binaural two-dimensional locations and recovered ear depths, the left-ear and right-ear positions can be jointly reconstructed in three-dimensional space. For each frame k, the binaural coordinates in the cabin frame are written as

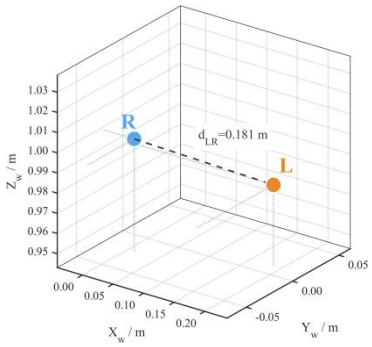
$$\mathbf{P}_w(k)=[\mathbf{p}_{w,L}(k)\mathbf{p}_{w,R}(k)]. \tag{7}$$

This joint form is preferred over two independent point outputs because the downstream control task requires a synchronized and structurally consistent binaural state rather than two unrelated position estimates (Figure 3).



(a) Binaural 2D localization in the original image

(b) Ear-region depth recovery in the corresponding frame



(c) Binaural 3D reconstruction in the cabin coordinate system

Figure 3 Representative Frame Showing Binaural 2D Localization, Ear-Region Depth Recovery, and Cabin-Coordinate 3D Reconstruction

3.5.2 Head-pose-based coordinate compensation

Although three-dimensional reconstruction already provides physical ear coordinates, the reconstructed positions are still affected by perspective variation and asymmetric visibility during head motion. To reduce pose-related systematic error, a head-pose compensation stage is introduced. The head pose at frame k is represented as

$$\boldsymbol{\eta}(k)=[\psi(k),\theta(k),\phi(k)]^{\mathrm{T}}, \tag{8}$$

where $\psi(k)$, $\theta(k)$, and $\phi(k)$ denote yaw, pitch, and roll, respectively. The raw reconstructed binaural coordinates are then corrected by a pose-related compensation term whose weight depends on observation validity. When the current observation quality is high, the system relies mainly on the frame-wise reconstruction result; when local mismatch or occlusion is detected, the pose-related compensation is given a larger role so as to suppress abrupt geometric inconsistency.

This treatment is especially important for yaw-dominant sequences, where self-occlusion and facial asymmetry can amplify reconstruction error in monocular side-view observations.

3.5.3 Temporal filtering and short-term prediction

After pose compensation, the binaural three-dimensional coordinates still remain time-varying observations rather than final control-usable states. Their frame-to-frame variation is influenced by depth noise, local landmark perturbation, and discrete visual sampling. To improve temporal consistency, the proposed method applies Kalman filtering to the binaural state sequence and further performs short-term prediction for output-time alignment.

For each ear, the filtered state evolves as

$$\hat{\mathbf{x}}_{k|k-1}=\mathbf{A}\hat{\mathbf{x}}_{k-1|k-1}, \tag{9}$$

$$\hat{\mathbf{x}}_{k|k}=\hat{\mathbf{x}}_{k|k-1}+\mathbf{K}_k\left(\mathbf{y}_k-\mathbf{H}\hat{\mathbf{x}}_{k|k-1}\right), \tag{10}$$

where $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k}$ are the predicted and posterior states, respectively, $\mathbf{y}_k$ is the observation vector, $\mathbf{A}$ is the state-transition matrix, $\mathbf{H}$ is the observation matrix, and $\mathbf{K}_k$ is the Kalman gain. Based on the posterior state, a short-horizon extrapolation is used to generate the final control-oriented binaural output at the required control instant.

As a result, the proposed method transforms frame-wise visual measurements into a continuous binaural ear-state sequence with improved smoothness, reduced jitter, and better interface usability (Figure 4).

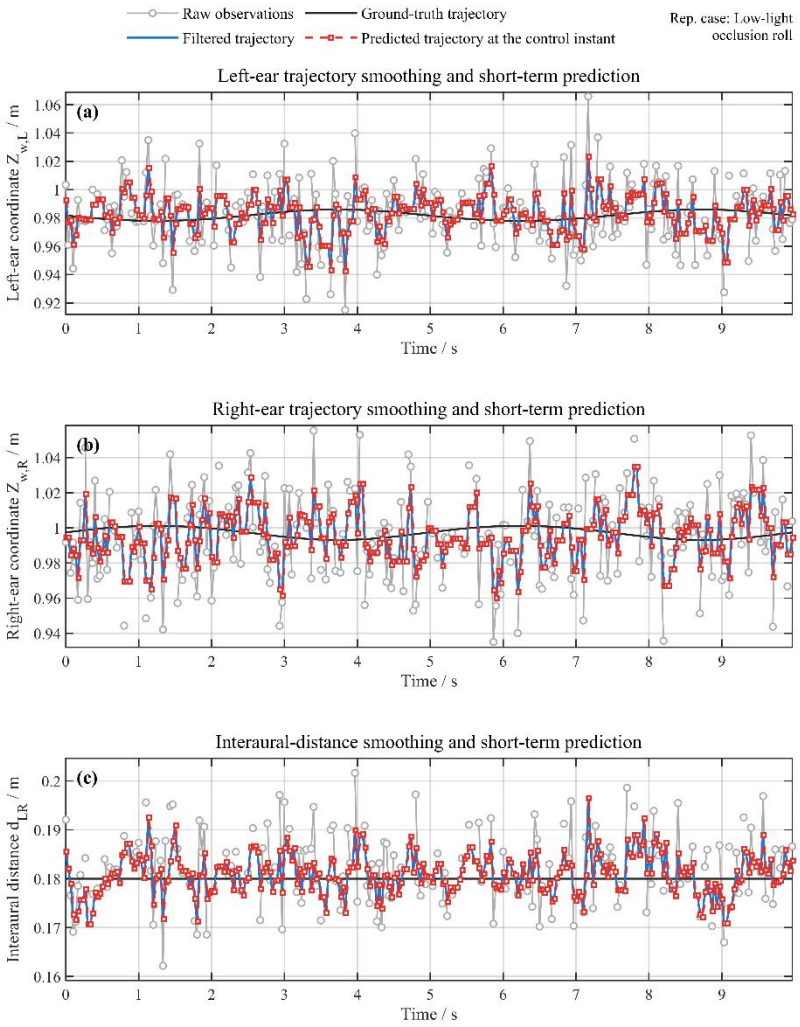


Figure 4 Temporal Smoothing and Short-Term Prediction of Binaural Trajectories in Continuous Video

4 EXPERIMENTAL SETUP AND EVALUATION METRICS

4.1 Experimental Platform and Test Conditions

Experiments were conducted on a simplified cabin visual platform with a fixed monocular camera installation, a seat-head region as the target observation area, and a consistent cabin-coordinate definition. To ensure comparability, the camera position, calibration parameters, and cabin coordinate system remained unchanged across all tests.

Six representative visual conditions were designed by combining illumination variation and local occlusion. In each condition, three head-motion sequences were recorded, namely pitch, roll, and yaw. The video length was fixed at 300 frames per sequence, and the visual update rate was 30 Hz (Table 1).

Table 1 Multi-Condition Test Settings

Case	Visual condition	Pose sequences
1	Daylight, unoccluded	Pitch, roll, and yaw
2	Daylight, partially occluded	Pitch, roll, and yaw
3	Low-light, unoccluded	Pitch, roll, and yaw
4	Low-light, partially occluded	Pitch, roll, and yaw
5	Fill-light, unoccluded	Pitch, roll, and yaw
6	Fill-light, partially occluded	Pitch, roll, and yaw

4.2 Evaluation Metrics

To evaluate both localization accuracy and temporal usability, five metrics were used.

The first is the binaural three-dimensional localization root-mean-square error (RMSE):

$$\mathrm{RMSE}_{3D}=\sqrt{\frac{1}{N}\sum_{k=1}^{N}\frac{\|\hat{\mathbf{p}}_{w,L}(k)-\mathbf{p}_{w,L}^{gt}(k)\|_2^2+\|\hat{\mathbf{p}}_{w,R}(k)-\mathbf{p}_{w,R}^{gt}(k)\|_2^2}{2}}. \tag{11}$$

The second is the mean three-dimensional error:

$$\bar{e}_{3D}=\frac{1}{N}\sum_{k=1}^{N}\frac{\|\hat{\mathbf{p}}_{w,L}(k)-\mathbf{p}_{w,L}^{gt}(k)\|_2+\|\hat{\mathbf{p}}_{w,R}(k)-\mathbf{p}_{w,R}^{gt}(k)\|_2}{2}. \tag{12}$$

The third is the mean interaural distance error:

$$\bar{e}_{LR}=\frac{1}{N}\sum_{k=1}^{N}\left|\|\hat{\mathbf{p}}_{w,L}(k)-\hat{\mathbf{p}}_{w,R}(k)\|_2-\|\mathbf{p}_{w,L}^{gt}(k)-\mathbf{p}_{w,R}^{gt}(k)\|_2\right|. \tag{13}$$

The fourth is the mean missing rate, which evaluates the proportion of frames in which a valid binaural observation is not directly available before temporal repair. The fifth is the maximum consecutive missing-frame count, which reflects the longest uninterrupted dropout interval in each condition.

Since the downstream control module uses filtered and predicted ear states rather than raw frame-wise observations, interface usability was also examined at the sequence level. In all conditions, the average interface latency of the final output sequence was 10 ms.

5 RESULTS AND DISCUSSION

5.1 Calibration Reliability

The geometric reliability of the proposed method depends first on camera calibration quality. The calibrated intrinsic parameters show that the effective focal lengths in the horizontal and vertical directions are highly consistent, while the principal point remains close to the image center. The two retained radial distortion coefficients also remain within a reasonable range for the adopted field of view and focal configuration. More importantly, the overall mean reprojection error is only 0.1535 px, indicating that the fitted monocular camera model accurately describes the actual imaging process and can support subsequent three-dimensional reconstruction.

The reprojection-error distribution across 30 calibration images further shows that the calibration quality is stable rather than relying on a few ideal samples. Although slight fluctuation exists among different checkerboard poses, no obviously abnormal calibration image dominates the final solution. This provides a reliable geometric basis for later ear-state reconstruction in the cabin coordinate system.

5.2 Overall Quantitative Performance under Six Conditions

The quantitative results of the six test conditions are summarized in Table 2.

Table 2 Quantitative Accuracy and Continuity Results under Six Conditions

Case	3D localization RMSE (mm)	Mean 3D error (mm)	Mean interaural distance error (mm)	Mean missing rate (%)	Maximum consecutive missing frames
1	18.40	16.58	0.37	0.00	0
2	19.48	17.43	1.20	15.44	6
3	20.85	18.50	1.97	0.00	0
4	25.57	22.32	3.33	23.67	6
5	18.70	16.82	0.83	0.00	0
6	21.25	19.10	1.83	13.95	4

Several clear observations can be made from Table 2. Under unoccluded conditions, the overall binaural three-dimensional localization error remains relatively low in Cases 1, 3, and 5, and no direct observation loss occurs. Once local occlusion is introduced, the mean three-dimensional error, interaural distance error, and missing-related indicators all increase. Among all conditions, Case 4, namely low-light with local occlusion, exhibits the largest error and the highest missing rate, indicating that illumination degradation coupled with partial occlusion is the most adverse scenario for monocular binaural tracking in the present cabin setup.

Nevertheless, even in these difficult conditions, the mean interaural distance error remains within the millimeter range. This shows that the relative geometric relationship between the left and right ears is preserved better than the absolute coordinate accuracy alone might suggest. For downstream ear-side active noise control, such structural consistency is especially important because the binaural state is used to describe not only two isolated points but also a physically meaningful head-related spatial configuration.

5.3 Influence of Head-Pose Type

The results further show that head-pose type is the most influential factor affecting localization accuracy. Averaged over all conditions, the mean three-dimensional errors for pitch, roll, and yaw sequences are 8.02 mm, 16.98 mm, and 30.38 mm, respectively, while the corresponding RMSE values are 9.64 mm, 18.46 mm, and 34.03 mm.
This trend is physically interpretable. In pitch-dominant motion, both ears remain comparatively visible in the image, and the main source of error comes from local fluctuation in depth estimation. Under roll motion, local perspective distortion becomes more obvious, and the error rises to the centimeter level. In yaw-dominant motion, the error further increases because monocular side-view observations are more sensitive to self-occlusion, facial asymmetry, and local depth bias near the lateral contour. These results indicate that yaw remains the most challenging motion pattern for monocular binaural reconstruction and is therefore the key target for later robustness improvement.

5.4 Effect of Illumination and Occlusion

A comparison among the six conditions further clarifies the robustness boundary of the proposed method. Under unoccluded conditions, the daylight case and the fill-light case show very similar overall performance, with mean three-dimensional errors of 16.58 mm and 16.82 mm, respectively. The low-light unoccluded case increases to 18.50 mm, which indicates that illumination degradation weakens local image contrast and reduces depth-boundary clarity, but the error level remains acceptable when no occlusion is introduced.
The role of fill light is also evident. After illumination assistance is added, the error level returns close to that of the daylight condition, showing that moderate fill light can effectively improve imaging stability in cabin scenes. By contrast, the effect of occlusion is more direct. The daylight occluded case increases to 17.43 mm, while the low-light occluded case reaches 22.32 mm. This confirms that local occlusion amplifies the difficulty of both landmark-based ear inference and ear-region depth recovery, especially when image contrast is already weakened by low illumination.
The fill-light occluded case still performs better than the low-light occluded case, with a mean three-dimensional error of 19.10 mm. This suggests that although fill light cannot completely remove the influence of occlusion, it can partially alleviate the degradation by stabilizing the visual front-end.

5.5 Trajectory Continuity and Interface Usability

From the viewpoint of engineering use, the main value of the proposed method does not lie only in frame-wise position accuracy, but also in whether the final ear-state sequence remains continuously usable for the control module. This issue is reflected by the missing rate, the maximum consecutive missing interval, and the final interface output behavior.
In Cases 1, 3, and 5, the missing rate is zero, indicating that direct frame-wise observations remain continuously available when no occlusion is introduced. In Cases 2, 4, and 6, however, the maximum consecutive missing interval reaches 6 frames in the most difficult cases. At a 30 Hz update rate, this corresponds to a raw visual interruption of approximately 0.20 s. If raw observations were directly fed into the control module, such discontinuity would severely affect downstream target updating. In the proposed framework, this problem is mitigated through pose compensation, Kalman filtering, and short-term prediction. As a result, although short-term visual failure still appears in difficult cases, the final ANC interface layer remains continuously available after temporal robustification.
This result is important because it highlights a distinction between raw vision performance and control-oriented usability. A monocular ear-tracking method for cabin active control should not be judged solely by frame-wise geometric error. Instead, it should be evaluated by whether it can provide a physically plausible and temporally continuous binaural state sequence for downstream control invocation. The present results show that the proposed framework satisfies this requirement under all six tested conditions.

5.6 Practical Implications and Applicability

Overall, the proposed method provides a practical tradeoff between deployment simplicity and control-oriented performance. Compared with depth-camera-based or infrared-based cabin perception systems, the monocular configuration is easier to integrate into existing intelligent cockpit architectures and imposes lower hardware cost. This advantage comes at the price of higher sensitivity to yaw motion, local occlusion, and weak-light imaging. However, the present results show that through the combination of geometric mapping, facial-structure-based ear inference, cabin-constrained depth recovery, and temporal robustification, the final binaural output can still maintain both acceptable accuracy and strong interface continuity.
Therefore, the main engineering contribution of the method is not to pursue the smallest possible frame-wise reconstruction error, but to establish a stable and low-cost perception front-end that can reliably support downstream ear -side active noise control in realistic cabin conditions.

6 CONCLUSION

This paper proposed a monocular-vision-based method for binaural three-dimensional localization and dynamic tracking in vehicle cabins, aiming to provide a control-oriented ear-state sequence for downstream ear-side active noise control. The proposed framework established a unified geometric mapping from image-plane observations to cabin-coordinate ear positions, combined facial-landmark-based binaural two-dimensional inference with cabin-feature-

constrained monocular depth estimation, and further improved temporal usability through head-pose compensation, Kalman filtering, and short-term prediction.

Experimental validation under six representative conditions demonstrated that the proposed method can maintain both localization accuracy and trajectory continuity under illumination variation, partial occlusion, and head-pose change. Across all tested cases, the root-mean-square error of binaural three-dimensional localization ranged from 18.40 to 25.57 mm, while the mean interaural distance error remained within 0.37 to 3.33 mm. The results also showed that yaw-dominant motion and the combined low-light-and-occlusion condition were the most challenging scenarios for monocular binaural reconstruction. Nevertheless, after temporal robustification, the final ear-state sequence remained physically plausible, geometrically consistent, and continuously available to the downstream control module.

Overall, the main contribution of this work lies not in pursuing the theoretical limit of frame-wise reconstruction accuracy, but in establishing a low-cost and practically deployable perception front-end that can reliably support ear-side active noise control in realistic cabin environments. Future work will focus on improving robustness under severe yaw motion and stronger occlusion, extending validation to more diverse occupants and cabin layouts, and integrating the proposed visual tracking module into a real-time closed-loop ear-side active noise control system.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

CONSENT FOR PUBLICATION

Written informed consent for publication of the identifiable facial images shown in Figures 2 and 3 was obtained from the participant, who authorized their use for academic publication.

REFERENCES

[1] Elliott S J, Nelson P A. Active noise control. IEEE Signal Processing Magazine, 1993, 10(4): 12-35.
[2] Jung W, Elliott S J, Cheer J. Combining the remote microphone technique with head-tracking for local active sound control. The Journal of the Acoustical Society of America, 2017, 142(1): 298-307.
[3] Elliott S J, Jung W, Cheer J. Head tracking extends local active control of broadband sound to higher frequencies. Scientific Reports, 2018, 8: 5403.
[4] Dong Y, Hu Z, Uchimura K, et al. Driver inattention monitoring system for intelligent vehicles: A review. IEEE Transactions on Intelligent Transportation Systems, 2010, 12(2): 596-614.
[5] Kaplan S, Guvensan M A, Yavuz A G, et al. Driver behavior analysis for safe driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(6): 3017-3032.
[6] Ji Q, Yang X. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. Real-Time Imaging, 2002, 8(5): 357-377.
[7] Bergasa L M, Nuevo J, Sotelo M A, et al. Real-time system for monitoring driver vigilance. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(1): 63-77.
[8] Mishra A, Lee S, Kim D, et al. In-cabin monitoring system for autonomous vehicles. Sensors, 2022, 22(12): 4360.
[9] Sharma P K, Chakraborty P. A review of driver gaze estimation and application in gaze behavior understanding. Engineering Applications of Artificial Intelligence, 2024, 133: 108117.
[10] Murphy-Chutorian E, Doshi A, Trivedi M M. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation//2007 IEEE Intelligent Transportation Systems Conference. IEEE, 2007: 709-714.
[11] Murphy-Chutorian E, Trivedi M M. Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(4): 607-626.
[12] Morency L P, Whitehill J, Movellan J. Monocular head pose estimation using generalized adaptive view-based appearance model. Image and Vision Computing, 2010, 28(5): 754-761.
[13] Alioua N, Amine A, Rogozan A, et al. Driver head pose estimation using efficient descriptor fusion. EURASIP Journal on Image and Video Processing, 2016, 2016(1): 2.
[14] Diaz-Chito K, Hernández-Sabaté A, López A M. A reduced feature set for driver head pose estimation. Applied Soft Computing, 2016, 45: 98-107.
[15] Wang Y, Yuan G, Mi Z, et al. Continuous driver's gaze zone estimation using RGB-D camera. Sensors, 2019, 19(6): 1287.
[16] Wang Y, Yuan G, Fu X. Driver's head pose and gaze zone estimation based on multi-zone templates registration and multi-frame point cloud fusion. Sensors, 2022, 22(9): 3154.
[17] Park B K D, Jones M, Miller C, et al. In-Vehicle Occupant Head Tracking Using a Low-Cost Depth Camera//WCX World Congress Experience. SAE Technical Paper, 2018.
[18] Tambwekar A, Park B K D, Kusari A, et al. Three-Dimensional Posture Estimation of Vehicle Occupants Using Depth and Infrared Images. Sensors, 2024, 24(17): 5530.
[19] Ko K L, Yoo J S, Han C W, et al. Pose and shape estimation of humans in vehicles. IEEE Transactions on Intelligent Transportation Systems, 2023, 25(1): 402-416.

[20] Cavalcanti U L, Poggi M, Tosi F, et al. CabNIR: A Benchmark for In-Vehicle Infrared Monocular Depth Estimation//2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025: 2578 -2590.
[21] Kuo S M, Morgan D R. Active noise control: a tutorial review. Proceedings of the IEEE, 2002, 87(6): 943-973.
[22] Pawelczyk M. Adaptive noise control algorithms for active headrest system. Control Engineering Practice, 2004, 12(9): 1101-1112.
[23] Elliott S J, Jones M. An active headrest for personal audio. The Journal of the Acoustical Society of America, 2006, 119(5): 2702-2709.
[24] Jiang H, Chen H, Tao J, et al. Accuracy requirements of ear-positioning for active control of road noise in a car. Applied Acoustics, 2024, 225: 110164.
[25] Liu Y, Li H, Zou H, et al. Active headrest combined with a depth camera-based ear-positioning system. The Journal of the Acoustical Society of America, 2025, 157(1): 519-526.
[26] Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.
[27] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1867-1874.
[28] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.
[29] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate O(n) solution to the PnP problem. International Journal of Computer Vision, 2009, 81(2): 155-166.
[30] Gui M, Schusterbauer J, Prestel U, et al. DepthFM: Fast monocular depth estimation with flow matching. arXiv preprint arXiv:2403.13788, 2024.