Non-Visual Human Eye Gaze Tracking Using Radar and Artificial Intelligence for Robust Driver Monitoring Systems

Master's thesis in Information and Communication Technology

Mengyuan Zhou, Yuqing Jiang

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Non-Visual Human Eye Gaze Tracking Using Radar and Artificial Intelligence for Robust Driver Monitoring Systems

Mengyuan Zhou, Yuqing Jiang

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Non-Visual Human Eye Gaze Tracking Using Radar and Artificial Intelligence for Robust Driver Monitoring Systems
Mengyuan Zhou, Yuqing Jiang

Supervisor: Frank Wang, Discernity AI
Examiner: Fredrik Brännström, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Non-Visual Human Eye Gaze Tracking Using Radar and Artificial Intelligence for Robust Driver Monitoring Systems
Mengyuan Zhou, Yuqing Jiang
Information and Communication Technology, Msc Progr
Chalmers University of Technology

Abstract

Estimating gaze direction from in-cabin sensors is central to driver distraction monitoring. Existing methods rely on cameras, which degrade under adverse lighting and collect identifiable facial imagery. Millimetre-wave radar avoids both limitations, and prior work at 60 GHz has demonstrated sensitivity to eye-region micro-motion such as blinks. Whether radar can support directional gaze estimation, rather than binary event detection, remains an open question.

This thesis develops a radar-only gaze estimation pipeline built on a 60 GHz FMCW sensor. Amplitude and inter-receiver phase-difference cues are extracted from the radar return, and candidate eye-movement events are detected from the radar signal alone, removing the need for camera or stimulus timing at inference. A lightweight dual-stream temporal architecture, DualStreamGazeNet, encodes the two cue types separately, fuses them through cross-modal self-attention, and produces both a direction label and a continuous gaze angle through a modality-aware hierarchical classifier and an auxiliary regression head.

Experiments on a multi-session dataset show that the model achieves 85.7% balanced accuracy for four-direction classification with azimuth and elevation errors of 7.82° and 3.15°. Under strict cross-session evaluation, few-shot calibration with three to five labelled events per direction yields 82.1% balanced accuracy, demonstrating that the learned representation generalises effectively with minimal target-session adaptation. These results establish that close-range mmWave radar is a viable modality for event-level gaze estimation and can serve as a privacy-preserving, illumination-independent complement to driver monitoring system.

Keywords: FMCW radar, Signal processing, Gaze tracking, Deep learning

Acknowledgements

We would like to express our sincere gratitude to everyone who contributed to the completion of this master's thesis at Chalmers University of Technology.
Our deepest thanks go to Discernity AI for the project opportunity, and especially to our supervisor, Frank Wang, for his professional guidance and support. We are equally grateful to our university advisor, Elina Amani, for her invaluable feedback and patience throughout the writing process, and to our examiner, Professor Fredrik Brännström, for his constructive guidance and oversight.
Finally, we thank Chalmers for offering an exceptional learning environment that made this work possible.

Yuqing Jiang & Mengyuan Zhou, Gothenburg, May 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

CNN	Convolutional Neural Network
FMCW	Frequency-Modulated Continuous-Wave
GT	Ground Truth (e.g. camera-based gaze labels)
HP	High-pass (filtering)
IF	Intermediate Frequency (beat signal after mixing)
IQ	In-phase and Quadrature (complex baseband components)
LP	Low-pass (filtering)
MTI	Moving Target Indication (slow-time clutter suppression)
RX	Receive (antenna chain or channel index)
STFT	Short-Time Fourier Transform

Nomenclature

Indices

i	Sample, radar-frame, or evaluation-sample index, depending on context
j	Camera-frame index
c	Class index
c_i	Discrete gaze-direction class label for sample i
e	Eye index, with $e \in \{L, R\}$
t	Time or frame index
τ	Radar slow-time index used for event proposal
f	Frequency index
r	Receive-antenna index
m	Feature-stream or modality index

Sets

$\{L, R\}$	Set of left and right eyes
$\{0, 1, 2\}$	Set of receive-antenna indices
$\{\text{LEFT}, \text{RIGHT}, \text{UP}, \text{DOWN}\}$	Set of gaze-direction classes
$\mathcal{B}$	Analysis frequency band used for rapid eye-motion energy
$\mathcal{I}$	Adaptive set of detected candidate event times
$\mathcal{W}$	Temporal window of consecutive radar frames

Parameters and Constants

T	Number of radar frames in an input window
T'	Temporal length after pooling in the neural encoder
T_b	Duration of the centre-fixation baseline interval
C_{feat}	Number of radar feature channels
C_{cls}	Number of gaze classes
d	Latent feature dimension
R_{eye}	Anatomical radius of the eyeball
r_0	Nominal range from the radar to the eye region
W_{eye}	Approximate horizontal eye width used for camera angle scaling
λ	Radar wavelength

λ_{reg}	Weight of the regression loss term
κ	Sensitivity parameter of the adaptive radar-event threshold
μ_τ	Local background mean of the radar motion-energy curve
σ_τ	Local background standard deviation of the radar motion-energy curve
K	Number of calibration events per gaze direction
N	Number of evaluated samples
W	Physical screen width
H	Physical screen height
D	Viewing distance from participant to screen
s_{arc}	Arc length along the anterior eye surface
s_{cam}	Camera ratio-to-angle scale factor

Variables and Functions

$z_r(t)$	Complex radar response for receive antenna r at frame t
$\mathbf{x}(t)$	Instantaneous six-channel radar feature vector at frame t
$a_r(t)$	Window-normalized amplitude of receive antenna r
$\Delta\phi_{10}(t), \Delta\phi_{20}(t)$	Inter-receiver phase-difference features relative to receiver 0
$\Delta a_{10}(t)$	Inter-receiver amplitude-difference feature $a_1(t) - a_0(t)$
X_i	Event-centred radar feature-window tensor for sample i
$X_i^{(a)}$	Amplitude-like stream input for sample i
$X_i^{(p)}$	Phase-stream input for sample i
$F_i^{(a)}, F_i^{(p)}, F_i^{(m)}$	Encoded feature maps for amplitude, phase, or modality stream m
E_a, E_p	Amplitude and phase temporal encoders
f_i	Pooled shared representation for sample i
$\mathbf{g}(t)$	Camera-derived gaze vector at time t
$\mathbf{g}_i$	Camera-derived gaze vector aligned to radar frame or sample i
$\hat{\mathbf{y}}_i^{\text{reg}}$	Two-dimensional regression output for sample i
θ_{az}	Azimuth gaze angle
θ_{el}	Elevation gaze angle
$\hat{\theta}_{\text{az},i}, \hat{\theta}_{\text{el},i}$	Predicted azimuth and elevation angles for sample i
$\mathbf{q}(t)$	Normalized iris position in the local eye coordinate frame
$\mathbf{q}_0$	Centre-fixation baseline of the normalized iris position
$\mathbf{S}$	Scale and sign mapping from relative image displacement to gaze angle
$x_{I,e}$	Horizontal iris-centre coordinate for eye e
$y_{B,e}$	Vertical coordinate of the bottom iris landmark for eye e
$x_{\min,e}, x_{\max,e}$	Horizontal eye-contour limits for eye e
$\bar{y}_{C,e}$	Mean vertical coordinate of the canthus points for eye e
$h_e(t), v_e(t)$	Horizontal and vertical iris ratios for eye e

$h(t), v(t)$	Binocular horizontal and vertical iris ratios
b_h, b_v	Horizontal and vertical centre-baseline ratios
$\Delta h(t), \Delta v(t)$	Relative horizontal and vertical gaze displacements in ratio space
$j^*(i)$	Camera-frame index matched to radar frame i
a_i, r_i, u_i	Axis, left-right branch, and up-down branch logits
$\log p_a, \log p_r, \log p_u$	Log-softmax probabilities for the hierarchical classifier heads
$\ell_{i,\mathrm{LEFT}}, \ell_{i,\mathrm{RIGHT}}, \ell_{i,\mathrm{UP}}, \ell_{i,\mathrm{DOWN}}$	Composed four-way direction logits for sample i
z_i	Hierarchical axis target for sample i
$\mathcal{L}$	Multi-task learning objective
$\mathcal{L}_i$	Per-sample multi-task training objective
$\mathcal{L}_{\mathrm{cls},i}$	Classification loss for sample i
$\mathcal{L}_{\mathrm{reg},i}$	Regression loss for sample i
$\mathcal{L}_{\mathrm{hier},i}$	Hierarchical classification loss for sample i
w_i	Sample weight derived from camera-estimated gaze displacement
$S(\tau, f)$	Time-frequency representation of the selected radar return
$E(\tau)$	Band-limited radar motion energy
M	Confusion matrix
BAcc	Balanced accuracy
$\mathrm{MAE}_{\mathrm{az}}, \mathrm{MAE}_{\mathrm{el}}$	Mean absolute errors for azimuth and elevation
$(x_i, y_i), (\hat{x}_i, \hat{y}_i)$	Known and predicted normalized screen-target coordinates
e_i	Screen-target angular error for target i
$t_{\mathrm{start}}, t_{\mathrm{end}}$	Start and end frame indices of a detected radar event
t_{c}	Anchor or centre frame of a radar-event window
$\mathbf{y}_i$	Continuous azimuth-elevation label vector for sample i

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background and Motivation

Driver monitoring systems are designed to observe driver attention continuously and reliably in support of road safety. A driver's gaze direction is one of the major direct indicators of visual engagement with the forward road scene, and sustained off-road gaze is a well-established predictor of distraction and collision risk. Recent European Union regulations require Advanced Driver Distraction Warning systems in all new vehicles, reflecting regulatory recognition that gaze monitoring is a safety-critical function [1].

Camera-based gaze estimation has been the dominant approach, with appearance-based deep learning methods achieving strong performance under constrained conditions [2], [3]. In practice, however, in-cabin cameras are sensitive to illumination changes, eyewear, and direct sunlight, and they capture identifiable facial imagery that raises privacy concerns [1].

Millimetre-wave radar offers a complementary alternative. At 60 GHz, the wavelength is approximately 5 mm, so sub-millimetre surface displacements produce measurable phase changes in the reflected signal, and the sensor operates independently of ambient light without forming facial images [4]. Recent work has shown that close-range radar can detect eye-blink events [5], characterise eyelid dynamics at millimetre-wave frequencies [6], and estimate eye rotation angle [7]. However, few studies have addressed four-direction gaze classification and continuous gaze angle regression from short event-centred radar windows recorded by a 60 GHz sensor, evaluated under leave-one-session-out conditions.

1.2 Aim

The aim of this thesis is to determine whether short-range 60 GHz Frequency-Modulated Continuous-Wave(FMCW) radar can support four-direction gaze event classification and continuous gaze angle regression from event-centred radar feature windows, and to develop and evaluate a multi-task neural architecture that separates and fuses amplitude and phase-difference information for this purpose.

1.3 Research Questions

The investigation is organised around three research questions, progressing from basic signal feasibility to cross-session robustness and radar-only operation:

- **Q1: Radar gaze information.** Do short event-centred 60 GHz FMCW radar windows contain learnable structure that supports both four-direction gaze classification and continuous azimuth and elevation regression?
- **Q2: Cross-session generalisation and calibration.** How well does a model trained on multiple recording sessions generalise to a held-out session, and to what extent can a small number of labelled target-session examples recover performance lost under strict leave-one-session-out evaluation?
- **Q3: Radar-only event localisation.** Can candidate eye-movement intervals be detected from a continuous radar recording without access to camera frames or stimulus timing, so that the trained gaze estimator can operate as a radar-only component at inference time?

1.4 Delimitations

The study is conducted in a controlled laboratory setting with a single 60 GHz radar module and a four-direction discrete stimulus protocol; in-vehicle validation, free-viewing gaze, and continuous point-of-regard estimation are outside the scope of this work. All evaluation protocols operate at the session level, and cross-subject generalisation is not evaluated. The radar-only event detection pipeline is assessed against camera-derived reference labels in controlled recordings but is not validated in continuous real-time operation. Camera frames and stimulus events are used for training supervision and offline evaluation and are not part of the inference path.

1.5 Contributions

This thesis makes three contributions. The first is a demonstration that 60 GHz close-range FMCW radar supports four-direction gaze event classification and continuous azimuth and elevation regression from short event-centred windows, with results reported under both leave-one-session-out and calibration-assisted evaluation protocols. The second is DualStreamGazeNet, a modality-aware dual-stream temporal architecture that encodes amplitude-like and phase-difference channels in separate convolutional encoders, fuses them through cross-modal attention, and produces direction predictions through a hierarchical classifier combined with an auxiliary regression head. The third is an empirical analysis showing that phase-difference features and a modality-aware hierarchical head are the principal contributors to session-robust performance, and that few-shot target-session calibration substantially closes the cross-session performance gap.

1.6 Thesis Outline

Chapter 2 covers FMCW radar fundamentals, eye movement kinematics, camera-based gaze ground-truth construction, machine learning for radar time series, and related work.
Chapter 3 describes the experimental system, data collection, radar feature extraction, model architecture, training objective, and evaluation protocols.
Chapter 4 reports classification, regression, ablation, and calibration results together with a radar-only inference demonstration.
Chapter 5 interprets the results in relation to the research questions and addresses limitations and future directions.
Chapter 6 summarises the principal findings.

2

Theory & Background

2.1 FMCW Radar Fundamentals

This work uses frequency-modulated continuous-wave (FMCW) radar near 60 GHz. A linear chirp sweeps bandwidth B over chirp duration T_c, giving chirp slope $\beta = B/T_c$ [8]. For a dominant scatterer at range R, the round-trip delay is $\tau = 2R/c$, where c denotes the speed of light.
Mixing the transmit and receive signals and retaining the low-pass term yields an approximately sinusoidal intermediate-frequency (IF) beat whose frequency scales linearly with the delay. The beat frequency f_b is given by

$$f_b = \beta\,\tau = \frac{2\beta R}{c} = \frac{2BR}{cT_c}\,. \tag{2.1}$$

Spectral analysis of the sampled IF signal maps beat frequency f_b to a range bin. Under ideal calibration, the estimated target range $\widehat{R}$ follows as

$$\widehat{R} = \frac{c\,f_b}{2\beta} = \frac{c\,f_b\,T_c}{2B}\,. \tag{2.2}$$

Two point targets separated by less than the Rayleigh range resolution, given by

$$\Delta R = \frac{c}{2B} \tag{2.3}$$

are not fully distinguished by magnitude-only range processing [8].
Phase is more sensitive to micromotion than magnitude bins. For a small radial displacement Δr at carrier wavelength $\lambda = c/f_0$, where f_0 is the carrier frequency, the resulting phase shift $\Delta\phi$ satisfies

$$\Delta\phi \approx \frac{4\pi\,\Delta r}{\lambda}\,, \tag{2.4}$$

which motivates exploiting slow-time phase and micro-Doppler structure in later chapters [9].

2.2 Radar Signal Processing for Micro-motion

Range information is obtained by windowed spectral estimation along fast time, followed by range bin selection. Along slow time (chirps or frames), micromotion

appears as structured amplitude variation and phase evolution. Inter-receiver phase differences suppress common-mode drift while retaining spatially differential motion [9].
Clutter and DC energy are attenuated with high-pass or moving-target-indication (MTI) style operators when needed [8]. Phase is unwrapped for analysis, and mild band-pass filtering can remove slow drift and out-of-band noise before correlation with oculomotor dynamics.
For time–frequency inspection, a short-time Fourier transform (STFT) is applied. With analysis window $w[n]$ of length N, hop size H, sample index n, frame index m, and frequency bin index k, the STFT spectrogram $S(m, k)$ of the discrete-time IF signal $x[n]$ is

$$S(m,k) = \sum_{n} x[n]\, w[n - mH]\, e^{-\mathrm{j}2\pi kn/N}\,, \tag{2.5}$$

which highlights micro-Doppler energy away from zero Doppler [9]. These steps support visual interpretation and event localisation, and learning features are constructed downstream without committing to a single estimator.

2.3 Eye Movement and Gaze Kinematics

Radar-based gaze estimation is possible only because gaze changes correspond to small but structured physical movements of the eye region. The adult eyeball is approximately spherical, with a typical diameter close to 24 mm [10]. A change in gaze direction therefore rotates the corneal and scleral surfaces relative to the radar, while eyelid motion can additionally alter how much of the anterior eye surface is exposed. At 60 GHz, these millimetre-scale changes are not negligible relative to the radar wavelength. The eye region is a compact but composite scattering target, consisting of the eyeball, eyelids, nearby skin, and small residual head motion.
The movements most relevant to the present study are saccades followed by short fixations. A saccade is a rapid, approximately ballistic rotation that moves the fovea from one target to another, and its amplitude, duration, and peak velocity follow the classical main-sequence relationship of human eye movements [11], [12]. Fixation is the subsequent gaze-holding interval, during which the eye is comparatively stable but not perfectly still. Blinks and partial eyelid closures are not gaze commands, yet they are important for radar sensing because they can produce strong amplitude changes by replacing the exposed corneal region with eyelid tissue. Eyelid kinematics are also coupled to vertical gaze, especially during upward and downward movements [13].
The geometric relation between eye motion and radar observables is summarized in Figure 2.1. In this simplified geometry, R_{eye} denotes the anatomical radius of the eyeball, and r_0 denotes the nominal range from the radar to the eye region. During an eye movement, the effective scattering point near the anterior eye surface is displaced slightly. The component of this local displacement along the radar line of sight is denoted by Δr, because this component changes the round-trip path length measured by the radar.
For a small eye rotation of angular magnitude θ, expressed in radians, the displacement along the curved anterior eye surface can be approximated by the arc length

$s_{\text{arc}} \simeq R_{\text{eye}}\theta$, where $R_{\text{eye}} \approx 12\,\text{mm}$. A 10° gaze rotation therefore corresponds to a surface displacement of approximately 2.1 mm. Only the line-of-sight projection of this motion contributes directly to the radar phase; denote this projected radial displacement by Δr. Under the small-displacement assumption used in Section 2.1, the associated phase excursion then follows (2.4), with carrier wavelength $\lambda = c/f_0$. With λ close to 5 mm at 60 GHz, even a radial displacement of 0.2 mm produces a phase change of approximately 0.5 rad. This order-of-magnitude argument does not imply that gaze angle can be recovered from a single phase sample, but it shows that saccade-scale eye motion can induce measurable phase perturbations when the eye-related return is sufficiently isolated from other motions.

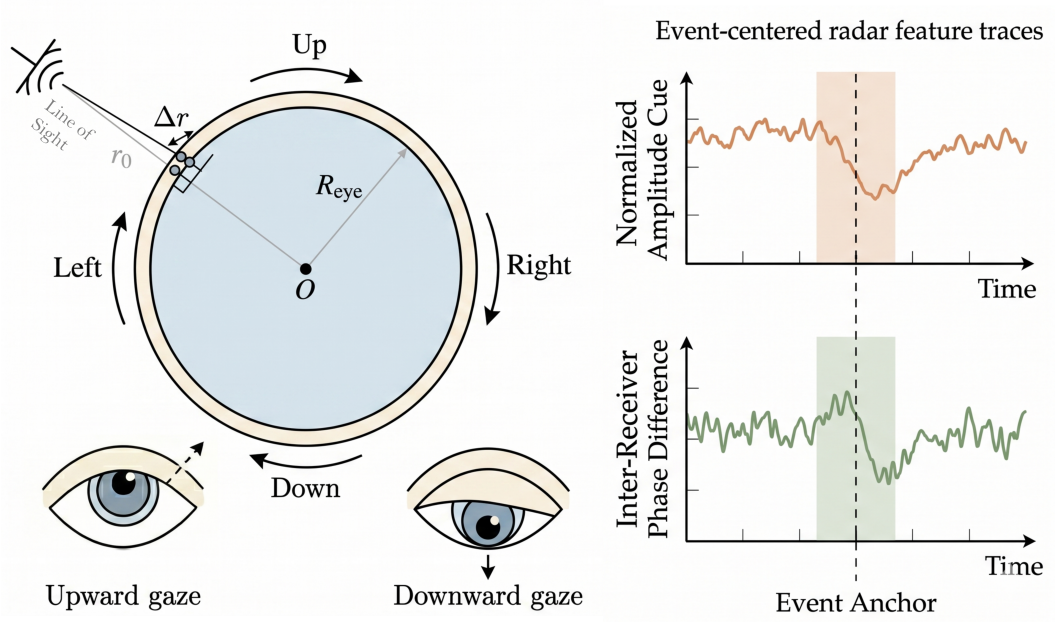


Figure 2.1: Conceptual relation between gaze kinematics and radar observables. Eye rotation changes the effective scattering point projected onto the radar line of sight, producing a small radial displacement component Δr.

The same mechanism also explains why gaze estimation from radar is sensitive to recording geometry. Small changes in head pose, subject position, and local scattering conditions can introduce path-length and amplitude variations of the same order as the eye motion itself. The useful information is therefore not expected to appear as a stable absolute phase or amplitude value at one instant. It is more naturally represented as a short temporal pattern around a gaze transition, including the pre-movement baseline, the saccade, and the early fixation response. Related FMCW radar work has similarly treated eye motion as a learnable radar event under external supervision [7], and millimetre-wave measurements of eyelid motion show that eyelid dynamics can contribute observable radar modulation [6].
Event-centred windows preserve the temporal structure of the gaze transitions, and the amplitude and phase cues capture complementary aspects of the radar response to eye motion. The sensitivity of the radar signature to recording geometry also motivates evaluation protocols that test how strongly the learned gaze cues depend on session-specific scattering conditions.

2.4 Camera-based Gaze Ground Truth

In a supervised radar gaze-estimation study, the eye movement that gives rise to the radar response must be observed by an independent reference modality. Camera-based gaze estimation provides this reference in the recorded experiments by converting visible eye motion into continuous azimuth and elevation labels. The camera is therefore used as a measurement instrument for constructing supervision and validation targets, while the gaze estimator itself is evaluated from radar-derived features.

Video-based gaze estimation commonly relies on geometric information from the eye region, such as pupil or iris location, eyelid contour, eye corners, and head pose. A long-standing difficulty is that these cues are affected by individual eye shape, partial occlusion, illumination, scale, and camera viewpoint [3]. In the present work, the camera reference is therefore treated as a practical supervision signal with known limitations. The controlled laboratory setup, frontal viewing geometry, and known stimulus timing make such a reference suitable for offline label generation, while its residual errors are still considered in the later quality analysis.

The camera pipeline used in this thesis follows the landmark-based paradigm. MediaPipe Face Mesh estimates a dense facial surface from monocular video, and the iris-refined configuration augments the face mesh with landmarks around the iris region [14], [15]. For gaze estimation, the central idea is to measure the iris position relative to stable local references of the eye, such as the eye corners and the eye contour. For small rotations, the displacement of the iris within this local eye coordinate frame can be interpreted as a proxy for horizontal and vertical gaze displacement. In abstract form, the camera-derived gaze vector can be written as

$$\mathbf{g}(t) = \begin{bmatrix} \theta_{\mathrm{az}}(t) \\ \theta_{\mathrm{el}}(t) \end{bmatrix} \approx \mathbf{S}\Big(\mathbf{q}(t) - \mathbf{q}_0\Big), \tag{2.6}$$

where $\mathbf{q}(t)$ denotes the normalized iris position in the local eye frame, $\mathbf{q}_0$ is the centre-fixation baseline for the same recording session, and $\mathbf{S}$ is a scale and sign convention that maps relative image displacement to azimuth and elevation. This expression is kept at the principle level. The detailed construction of the horizontal and vertical ratios, the session baseline, and the alignment to radar frames is specified in the Methods chapter.

The camera reference and the discrete stimulus labels serve different supervisory roles. The camera-derived angles provide continuous gaze displacement for regression, while the direction labels are taken from the stimulus protocol, where the intended target direction is known. Using the controlled stimulus for classification avoids introducing class noise from thresholding a monocular gaze estimate. The sign convention and the detailed label construction are defined in Section 3.3.

Because the camera reference is estimated from monocular facial and iris landmarks, it remains subject to noise and bias. Landmark jitter, partial eyelid occlusion, small head-pose changes, and imperfect centre fixation can all perturb the inferred angles. These limitations motivate the empirical ground-truth quality checks reported in the Results chapter, including frame-level label availability, direction-wise angle consistency, and screen-target validation. They also motivate the use of robust regression

losses and separate azimuth and elevation error reporting in the radar experiments.

2.5 Machine Learning for Radar Time Series

The radar representation used for gaze estimation is naturally a multichannel time series. Each sample contains a short temporal window around an eye-movement event, and each channel describes one aspect of the radar return, such as reflected amplitude, receiver-dependent amplitude balance, or inter-receiver phase difference. The learning problem is therefore not only to classify an isolated feature vector, but to interpret a local temporal pattern whose informative portion may occupy only part of the window. This makes sequence modelling central to the radar gaze-estimation problem.

One-dimensional convolution is a common starting point for fixed-length time series because it learns local temporal filters with relatively few parameters. A convolutional encoder can detect short patterns such as a rapid transition, a post-saccadic settling response, or a transient amplitude change, while sharing the same filters across the window. Compared with fully connected models applied to flattened windows, temporal convolution preserves neighbouring-frame structure and reduces the number of trainable parameters. This property is useful when the number of training examples is limited and the discriminative signal is concentrated in short temporal intervals. The broader suitability of convolutional architectures for sequence modelling has been shown in empirical comparisons of temporal convolutional and recurrent networks [16].

Radar gaze features are also heterogeneous. Amplitude-related channels and phase-difference channels are measured from the same physical event, but they do not have identical statistical properties or the same geometric interpretation. Amplitude cues are affected by reflected signal strength, eyelid exposure, and receiver-dependent energy balance, whereas phase cues encode relative path-length changes across receiver pairs. A useful model therefore needs both modality-specific processing and later interaction between modalities. Attention mechanisms provide one way to perform this interaction, because they allow the model to weight different temporal positions and feature streams according to their relevance to the prediction. The self-attention mechanism introduced in the Transformer architecture formalizes this idea by computing data-dependent interactions between tokens within a sequence [17]. In time-series sensing, this principle can support fusing heterogeneous feature representations after each has first been encoded according to its own signal characteristics.

The target structure is also multi-faceted. A stimulus protocol defines a discrete gaze direction, while a camera reference provides a continuous azimuth and elevation estimate. These two forms of supervision are related but not interchangeable. Direction classification encourages separability among target classes, whereas angle regression encourages the learned representation to preserve continuous displacement information. This matches the general principle of multi-task learning, in which related tasks share intermediate representations while retaining task-specific output layers [18]. A compact way to express a combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(\hat{c}, c) + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}(\hat{\mathbf{g}}, \mathbf{g}), \tag{2.7}$$

where c is the discrete class label, $\mathbf{g}$ is the continuous angle target, and λ_{reg} controls the relative influence of the regression term. Robust losses such as the Huber or smooth L_1 loss are useful for the regression term when the continuous labels may contain occasional noise or outliers [19].

The four gaze classes also have a simple semantic structure: left and right are horizontal movements, while up and down are vertical movements. Treating these classes as unrelated categories discards this structure. Hierarchical classification instead decomposes a label into decisions at multiple levels [20]. This decomposition is physically meaningful because horizontal and vertical eye movements need not produce identical radar signatures. The hierarchy does not change the external goal of predicting one of four directions, but it can encourage a representation that reflects the geometry of the task.

A final issue is domain shift across recording sessions. Radar micro-motion features are strongly affected by subject posture, sensor placement, local scattering geometry, and phase behaviour. A model that performs well when training and validation samples come from the same session may therefore degrade when evaluated on a held-out session. Domain-adversarial learning is one general strategy for reducing domain-specific information in learned representations [21], while few-shot calibration offers a more direct way to estimate the session-specific response of a sensing system. Both approaches are relevant when the number of recording sessions is small and session-level variation is a primary concern. Evaluation in such settings should therefore distinguish within-session learnability from cross-session generalization, because a single aggregate score can conflate the two.

2.6 Related Work

Camera-based gaze estimation constitutes the primary body of prior work against which the present approach is positioned. Since the early 2000s, the field has progressed from geometric models of the pupil and corneal reflection to appearance-based deep learning methods that operate from a single monocular camera without specialised illumination [3]. Recent surveys document how convolutional and transformer architectures have substantially improved cross-person accuracy and reduced the dependence on personal calibration [2]. For driver monitoring specifically, camera-based systems remain the commercial standard, but reviews of deployed systems consistently identify illumination variability, occlusion, and in-cabin privacy as the main factors limiting reliability and acceptance [1]. The camera-based reference pipeline used in Section 2.4 inherits both the capabilities and the limitations of this line of work; the goal of the present thesis is to investigate whether radar can provide gaze information independently of those limitations.

A smaller body of work has established that millimetre-wave radar can detect and characterise eye-region motion at close range. Patscheider et al. used a 120 GHz radar to measure eyelid dynamics and showed that the opening and closing of the eyelid produces measurable changes in the amplitude of the reflected signal [6].

Contactless detection of both eyelid blink and gross eyeball rotation at 120 GHz has further demonstrated that short-range radar can distinguish between different eye activity states [22]. These results confirm the physical observability discussed in Section 2.3, but they address eye activity classification rather than the estimation of gaze direction from learned temporal features.

In the driver monitoring context, radar has been applied to related but distinct problems. Hu et al. used an impulse-radio ultra-wideband radar to detect individual eye-blink events in a simulated driving setting, demonstrating robust performance under conditions of vehicular motion and reporting a median detection accuracy above 90% [5]. Xu et al. inferred driver attention from the direction and angle of head rotation measured with ultra-wideband radar, without requiring access to eye-level signals [23]. Jang et al. explored gaze estimation at 60 GHz by first estimating head orientation from the radar return and then using head pose as a proxy for gaze direction [24]. Each of these systems shares the driver monitoring motivation of the present work, but none estimates gaze direction from direct eye-level radar features extracted around saccadic events.

The most directly related prior work is the RadEye system of Zhang et al., which uses a sub-6 GHz FMCW radar to estimate continuous eye rotation angles from a distance [7]. RadEye demonstrates that radar-derived amplitude and phase features carry direction-specific information about voluntary eye movements, which is the central physical premise of the present thesis. A recent system in the wearable domain places sub-centimetre millimetre-wave radar sensors on a glasses frame, capturing eye reflections at millimetre-scale standoff to support eye tracking on smart glasses [25].

The present work differs from these predecessors in three respects. First, it uses a 60 GHz sensor in a stationary close-range configuration relevant to in-cabin driver monitoring, rather than a sub-6 GHz setup or a near-eye wearable. Second, it formulates the task as four-direction event classification combined with continuous azimuth and elevation regression, both supervised from short event-centred windows extracted around saccade-anchored intervals; RadEye estimates a continuous rotation angle without a discrete direction protocol, and the wearable system targets a glasses form factor without a session-level generalisation evaluation. Third, performance is evaluated under a leave-one-session-out protocol designed to quantify cross-session generalisation, and under a calibration-assisted variant that assesses how few labelled target-session examples are needed to recover performance. These distinctions motivate the modality-aware dual-stream architecture, hierarchical classifier, and evaluation protocols described in Chapter 3.

3

Methods

3.1 System Overview and Hardware Setup

Figure 3.1 summarizes the complete workflow used in this thesis. The system is organized around an event-level radar gaze-estimation pipeline. During data collection, a 60 GHz FMCW radar observes the subject's eye and face region while a camera and a screen-based arrow stimulus are recorded on the same session timeline.

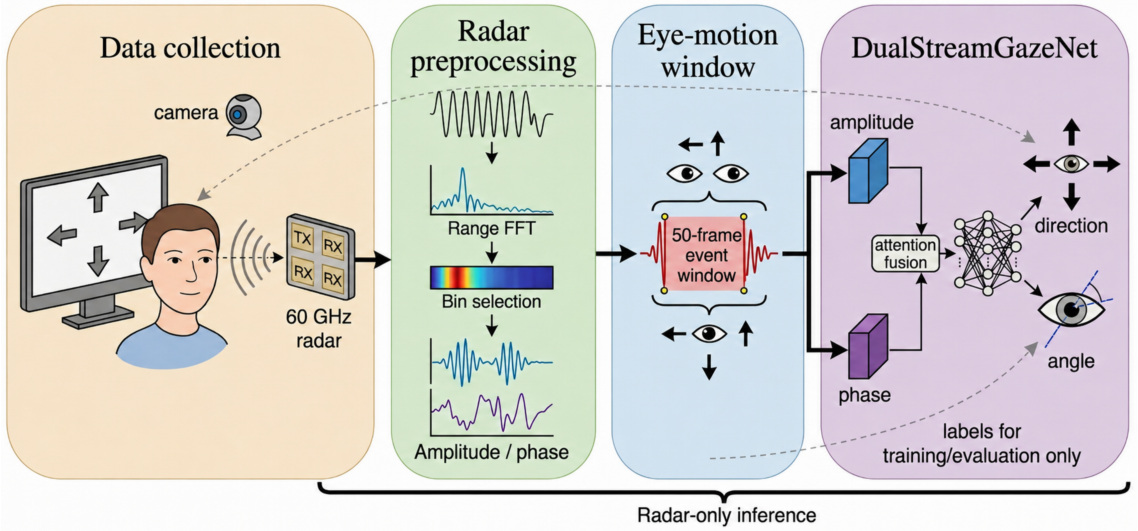


Figure 3.1: System overview of the radar-based gaze-estimation framework

The radar branch first converts the raw FMCW recordings into gaze-related feature sequences. After range-domain processing and range-bin selection, the selected radar return is represented through amplitude-related and phase-related cues. These features are chosen because small gaze movements can affect both the reflected signal strength and the relative phase behaviour across receiver channels. The resulting representation is then converted into a fixed-length event-centred window. In the current dataset, each sample contains a 50-frame temporal window around a gaze event, preserving the short transition and early fixation response that are expected to carry most of the discriminative information.
The reference branch assigns targets to these radar windows. The screen stimulus defines the discrete direction class, and the camera stream is processed offline to estimate continuous azimuth and elevation angles. The stimulus provides a stable classification target, while the camera provides a continuous regression target

and an evaluation reference. This separation avoids deriving direction classes from noisy camera-angle thresholds while still allowing the model to learn angular gaze displacement.
The estimator used in this work is DualStreamGazeNet. It separates amplitude-like channels and phase-difference channels into two temporal streams, fuses the encoded representations with attention, and produces both a direction prediction and a continuous azimuth/elevation estimate from the shared representation.
At test time, the intended deployment path is radar-only. A radar recording or stream is processed to detect candidate eye-movement intervals, each interval is converted into the same radar feature-window format used during training, and the trained model predicts gaze direction probabilities together with continuous angle estimates. Camera frames and stimulus events are only used to construct labels and compute offline metrics. The following sections describe each component of this overview in more detail.

3.1.1 Radar Front-end

Millimetre-wave sensing is performed with an Infineon BGT60TR13C FMCW radar module operated through the vendor FMCW driver stack. The integrated antenna layout shown in Figure 3.2 provides multiple spatial phase centres along one axis, supporting interferometric phase comparisons across receivers, a prerequisite for the phase-difference features used later in this thesis.

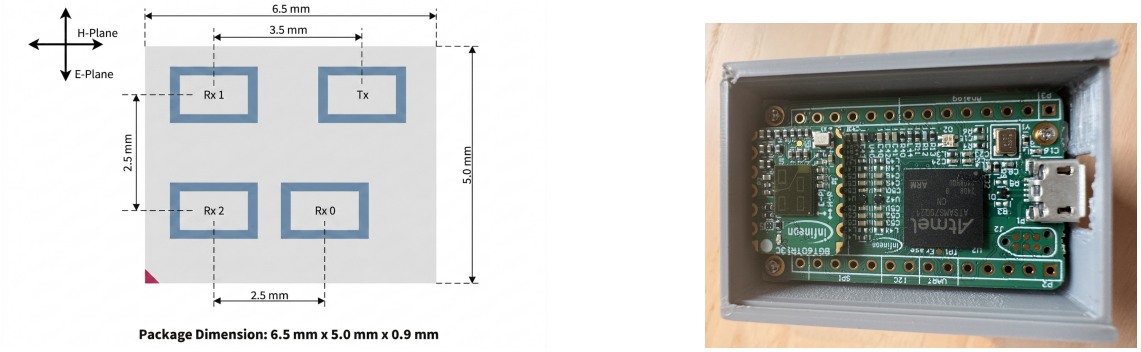


Figure 3.2: Antenna layout of the BGT60TR13C FMCW radar and the sensor assembly used in the laboratory recordings.

3.1.2 Radar Acquisition Parameters

Table 3.1 summarises the chirp and frame configuration programmed at acquisition time.
From swept bandwidth B, the coherent range resolution $\Delta R_{\mathrm{coh}} = c/(2B) \approx 2.7\,\mathrm{cm}$, where c is the speed of light. Realised resolution cells also depend on windowing and spectral leakage in downstream processing. The sampled chirp duration satisfies $N_{\mathrm{s}}/f_{\mathrm{s,IF}} = 0.8\,\mathrm{ms}$, where sample count N_{s} and IF sample rate $f_{\mathrm{s,IF}}$ are taken from Table 3.1. This is consistent with fast-time acquisition that is independent

Table 3.1: FMCW radar configuration used during data collection.

Quantity	Value
Start / end sweep frequency	58.0 GHz / 63.5 GHz
Effective swept bandwidth B	5.5 GHz
Samples per chirp N_s	400
IF sample rate $f_{\mathrm{s,IF}}$	500 kHz
Chirps per frame N_c	16
Chirp repetition interval	1 ms
Frame repetition interval	20 ms (50 frames/s)

of the programmed 1 ms chirp repetition interval. Figure 3.3 illustrates the resulting hierarchical timing structure. Each frame period of 20 ms contains $N_\mathrm{c} = 16$ chirps that occupy the first 16 ms (80 %), followed by a 4 ms inter-frame idle gap (20 %). Within each chirp of duration $T_\mathrm{chirp} = 1$ ms, the first 0.8 ms constitutes the IF sampling window during which $N_\mathrm{s} = 400$ samples are acquired at sample rate $f_\mathrm{s} = 500$ kHz, and the remaining 0.2 ms serves as an idle guard interval.

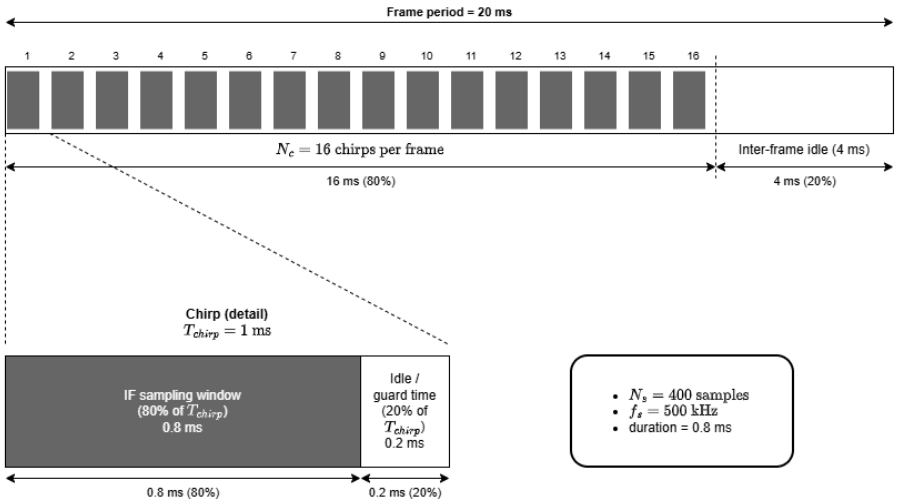


Figure 3.3: Hierarchical timing of chirps and frames under the configured FMCW sequence. Each 20 ms frame contains $N_\mathrm{c} = 16$ chirps over 16 ms, followed by a 4 ms inter-frame idle gap. Each chirp of duration $T_\mathrm{chirp} = 1$ ms is partitioned into an 0.8 ms IF sampling window ($N_\mathrm{s} = 400$ samples at $f_\mathrm{s} = 500$ kHz) and a 0.2 ms idle guard interval.

3.1.3 Camera and Computing Platform

Colour imagery for gaze supervision is captured with an Intel RealSense RGB-D sensor at 640×480 resolution and 30 fps. Processing and recording run on a laboratory workstation under Windows. Radar frames are buffered through bounded queues and written asynchronously to disk to avoid blocking real-time acquisition threads.

3.1.4 Participant, Geometry, and Stimulus Display

Participants were seated at a distance of 15 cm from the radar, with their chin supported by a rest to minimise spontaneous head movement during recording. As shown in Figure 3.4, the radar and camera module are mounted in front of the participant. Visual stimuli are rendered fullscreen on a 1920×1080 display.

Figure 3.4: Laboratory geometry for synchronous radar–camera–stimulus acquisition.

3.2 Data Acquisition and Synchronization

3.2.1 Concurrent Streams and Session Packaging

Each recording session produces a self-contained directory containing radar tensors and timestamps, camera imagery with per-frame annotations, monotonic host timestamps for alignment, and structured stimulus logs.

3.2.2 Radar Recording Path

The radar worker polls frames from hardware at the configured frame rate. The host records nanosecond-level timestamps, and frames are placed in the queue of the write thread, which stacks tensors together with paired timestamp arrays to form a continuous dataset in HDF5.

3.2.3 Stimulus Logging

The visual paradigm executes on the application main thread. Trials emit timed events including baseline onset, pseudo-random centre-hold intervals, outbound cues, peripheral fixation holds, and return cues. Each record carries a monotonic timestamp compatible with the radar and camera clocks.

3.2.4 Temporal Alignment Strategy

Host clocks are monotonic but not identical across devices, and alignment for offline processing therefore relies on matching event centres and sampling instants

rather than asserting sample-clock equality. Stimulus markers anchor gaze-related behaviours to protocol phases. Figure 3.5 shows that radar frames carry timestamps enclosing the hardware read window, and association uses nearest-neighbour mapping in post-processing with tolerance derived from half the radar frame period (10 ms). The two independent streams run at 30 Hz for the camera and 50 Hz for the radar, and alignment pairs each camera sample or gaze estimate with the temporally closest radar frame index. After acquisition, an offline routine may refine these intervals and export gaze angles.

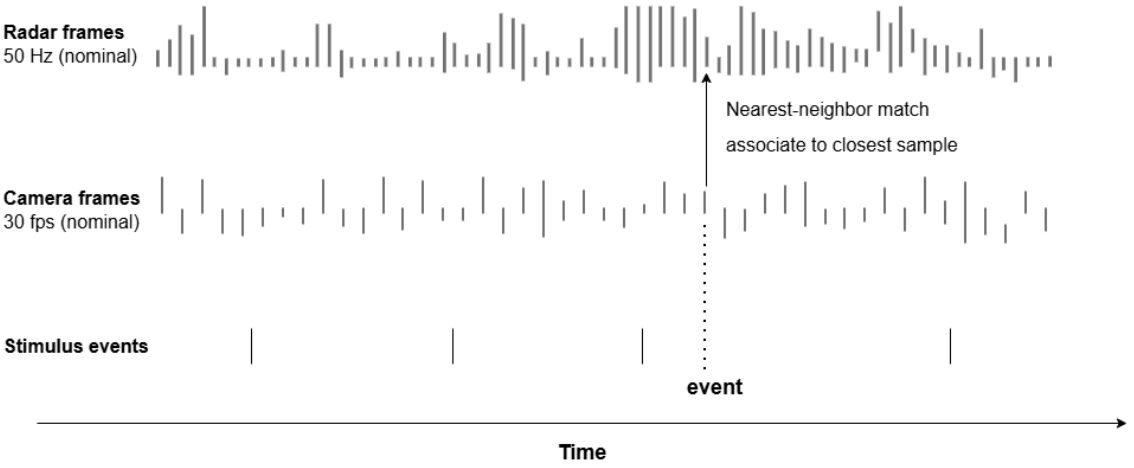


Figure 3.5: Illustration of asynchronous streams and offline temporal association (not to scale).

3.3 Camera-based Gaze Ground Truth

3.3.1 Reference Modality and Landmark Geometry

The camera stream provides the continuous gaze-angle reference used for supervised training and offline evaluation. During each recording session, RGB camera frames, radar frames, and timestamps were recorded on a common acquisition timeline. The camera frames were reprocessed offline with MediaPipe Face Mesh using iris-refined landmarks. MediaPipe Face Mesh estimates a dense facial surface from monocular video [14], and the iris-refined pipeline adds landmarks around the iris region [15]. The camera is therefore used as a reference modality during label construction, while the trained model receives only radar-derived feature windows during radar-only inference.

For each valid camera frame, the iris landmarks, eye-contour landmarks, and medial and lateral canthus points define a local eye coordinate system. Frames with missing or invalid landmarks are excluded from the ground-truth stream. For eye $e \in \{L, R\}$, let $x_{I,e}$ denote the horizontal coordinate of the iris centre, $x_{\min,e}$ and $x_{\max,e}$ the horizontal eye-contour extent, and $\bar{y}_{C,e}$ the mean vertical coordinate of the two canthus points. The horizontal iris ratio uses the iris centre, while the vertical ratio uses the bottom iris landmark $y_{B,e}$ rather than the iris-centre ordinate. This choice

avoids the direction-reversal artefact that occurs during upward gaze, when the upper eyelid partially occludes the iris and biases the estimated centre downward; the bottom landmark remains visible in all gaze directions and produces a monotonic vertical signal. The ratios are

$$h_e(t) = \frac{x_{I,e}(t) - x_{\min,e}(t)}{x_{\max,e}(t) - x_{\min,e}(t)}, \qquad v_e(t) = \frac{y_{B,e}(t) - \bar{y}_{C,e}(t)}{x_{\max,e}(t) - x_{\min,e}(t)}. \tag{3.1}$$

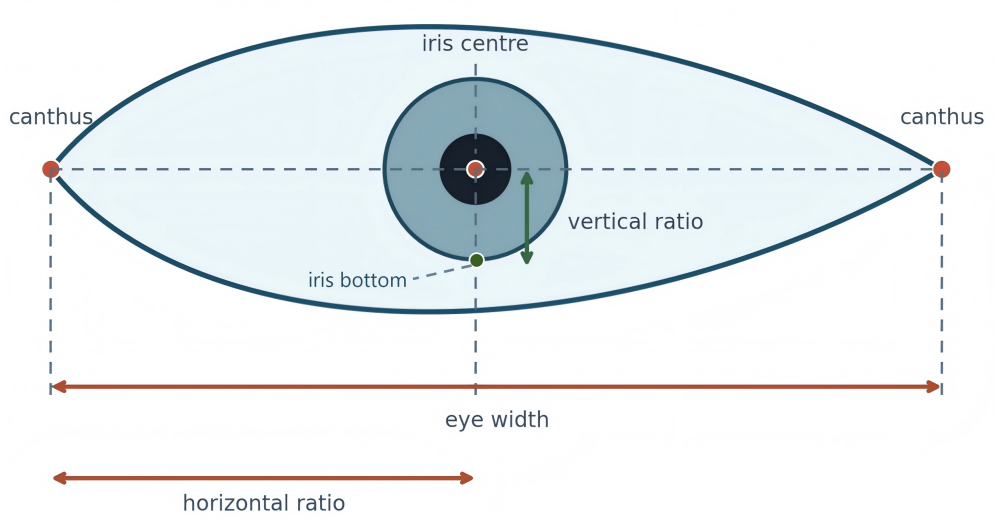


Figure 3.6: Schematic landmark geometry used for canthus-anchored iris-ratio computation. The figure identifies the iris centre, the canthus reference points, the eye-contour width, and the horizontal and vertical ratio axes.

Figure 3.6 illustrates the landmark geometry used in this computation. The binocular ratio is obtained by averaging the two eyes when both are valid, with the available eye used as a fallback when only one eye is valid. The vertical ratio is referenced to the canthus line and normalized by horizontal eye width rather than by eyelid opening. This design makes the vertical reference insensitive to blink state, squinting, and eyelid motion, which would otherwise affect both the vertical scale and the direction of the estimated signal.

3.3.2 Baseline Calibration and Angle Convention

A per-session centre baseline is estimated from the initial fixation interval, during which the subject looks at the central screen marker for $T_b = 5\,\mathrm{s}$. The baseline is computed as the median of all valid binocular ratios in this interval,

$$b_h = \mathrm{median}_{t<T_b}\, h(t), \qquad b_v = \mathrm{median}_{t<T_b}\, v(t), \tag{3.2}$$

where $h(t)$ and $v(t)$ denote the binocular averages. Relative gaze displacement is then defined as $\Delta h(t) = h(t) - b_h$ and $\Delta v(t) = v(t) - b_v$. The median baseline compensates for subject-specific eye geometry, camera placement, and small neutral-position offsets without requiring a full multi-point camera calibration.

The relative iris ratios are converted to angular gaze estimates using a small-rotation anatomical approximation. Adult eyeball diameter is commonly close to 24 mm, giving $R_{\text{eye}} \approx 12\,\text{mm}$ [10]. The horizontal palpebral fissure width is commonly around 27 to 29 mm [26]; using $W_{\text{eye}} \approx 28\,\text{mm}$ gives

$$s = \frac{W_{\text{eye}}}{R_{\text{eye}}} \approx \frac{28}{12} = 2.33 \text{ rad.} \tag{3.3}$$

The camera-based azimuth and elevation angles are therefore defined as

$$\theta_{\text{az}}(t) = -s\,\Delta h(t), \qquad \theta_{\text{el}}(t) = s\,\Delta v(t). \tag{3.4}$$

A positive azimuth denotes rightward gaze and a negative azimuth denotes leftward gaze. A positive elevation denotes downward gaze and a negative elevation denotes upward gaze. The negative sign in the azimuth definition accounts for the image-coordinate convention, since an iris shift toward image right corresponds to a gaze rotation toward the subject's left. No temporal smoothing is applied to the camera-derived angles, so that saccade timing is not shifted before alignment to the radar stream.

3.3.3 Temporal Alignment and Label Assignment

Camera-frame angles are aligned to the radar time base by nearest-neighbour timestamp matching. For radar frame i, the corresponding camera frame is

$$j^*(i) = \arg\min_j \left| t_i^{\text{radar}} - t_j^{\text{cam}} \right|. \tag{3.5}$$

The aligned gaze vector for the radar frame is $\mathbf{g}_i = [\theta_{\text{az}}(t^{\text{cam}}_{j^*(i)}), \theta_{\text{el}}(t^{\text{cam}}_{j^*(i)})]^\top$. In the event-anchored dataset, the default regression label is sampled 20 radar frames after the onset of the directional stimulus, corresponding to the target-hold portion of the stimulus sequence. When the camera-velocity quality filter is enabled, the label frame can instead be selected from the detected end of the saccade.

The continuous regression target and the discrete classification target serve different purposes. The regression target is the aligned camera angle, which preserves horizontal and vertical angular variation. The direction-classification target is taken from the controlled stimulus direction rather than from thresholded camera angles, avoiding additional class noise near the centre and near axis boundaries. The camera-based ground truth should therefore be interpreted as a practical supervision and evaluation reference, not as an error-free measurement of gaze. Its reliability depends on face and iris visibility, lighting, camera placement, head stability, and the anatomical approximation used for angle conversion. These properties are evaluated empirically in Section 4.3.

3.4 Radar Feature Extraction and Dataset Construction

This study constructs learning-oriented radar features on the frame timeline of the FMCW recording. For each frame and each receive channel, the raw chirp-by-fast-time cube is transformed to the range domain using standard FMCW spectral processing per chirp. The complex responses at a given range bin are then averaged across all chirps within the same frame. The result is one complex sample $z_r(t) \in \mathbb{C}$ per frame index t and per receive antenna $r \in \{0, 1, 2\}$ at the selected bin, aligned in time with the frame-synchronous gaze supervision while suppressing within-frame fast-time fluctuations that are not resolved at the radar frame rate.

3.4.1 Signal Representation

For each receive antenna, the per-frame complex response at a candidate range bin is obtained from the chirp-averaged range profile described above. A session-wise target bin is selected from temporal variability of the magnitude profile rather than from a single static peak, so that subsequent modelling emphasises bins where slow-time dynamics, rather than fixed near-field leakage, are strongest.
Let t index radar frames after alignment to the master clock shared with gaze estimates. All six channels below are evaluated at each frame index t on the chirp-averaged complex samples $\{z_0(t), z_1(t), z_2(t)\}$ at the operating range gate.

3.4.2 Six-channel Feature Design

For each temporal window $\mathcal{W}$ of consecutive frames, the feature vector $\mathbf{x}(t)$ is a six-channel vector (Table 3.2) formed as

$$\mathbf{x}(t) = \Big[a_0(t),\, a_1(t),\, a_2(t),\, \Delta\phi_{10}(t),\, \Delta\phi_{20}(t),\, \Delta a_{10}(t)\Big], \tag{3.6}$$

Table 3.2: Six-channel radar feature definition (single selected range bin; frame axis).

Channel index	Name	Description
0	`amp_rx0_norm`	RX0 magnitude, window-normalised
1	`amp_rx1_norm`	RX1 magnitude, window-normalised
2	`amp_rx2_norm`	RX2 magnitude, window-normalised
3	`dphi_rx1_rx0`	Inter-RX phase (RX1 vs. RX0) from $z_1 z_0^*$
4	`dphi_rx2_rx0`	Inter-RX phase (RX2 vs. RX0) from $z_2 z_0^*$
5	`damp_rx1_rx0`	Amplitude difference $a_1 - a_0$

where the first three entries are normalised receive amplitudes $a_0(t)$, $a_1(t)$, $a_2(t)$, the next two are inter-receiver phase differences $\Delta\phi_{10}(t)$ and $\Delta\phi_{20}(t)$, and the sixth is amplitude contrast $\Delta a_{10}(t)$. Window-normalised amplitude $a_r(t)$ for receive antenna

$r \in \{0, 1, 2\}$ is

$$a_r(t) = \frac{|z_r(t)|}{\frac{1}{|\mathcal{W}|} \sum_{t' \in \mathcal{W}} |z_r(t')|}. \tag{3.7}$$

Inter-receiver phase differences $\Delta\phi_{10}(t)$ and $\Delta\phi_{20}(t)$ are computed modulo 2π from the complex cross-channel product (equivalently, from the phase of $z_r(t)\, z_0^*(t)$, where $z_0^*(t)$ denotes the complex conjugate of $z_0(t)$), so that common-mode phase drift largely cancels before temporal detrending. Amplitude contrast $\Delta a_{10}(t)$ is defined as

$$\Delta a_{10}(t) = a_1(t) - a_0(t). \tag{3.8}$$

In practice, narrowband Butterworth filtering may be applied to the differential-phase tracks inside each window to suppress slow baseline wander while preserving saccade-scale excursions. Amplitude channels use the same window $\mathcal{W}$ for normalisation so that relative level changes remain comparable across sessions.

3.4.3 Design Rationale

The first three channels retain normalised slow-time amplitude at each individual receiver rather than forming an inter-receiver amplitude difference. Unlike phase, reflected amplitude is not dominated by sub-millimetre path-length variations, so single-receiver amplitude remains a stable and informative cue. It is most useful for vertical gaze changes, where eyelid-related modulation produces distinct energy signatures that differ across receivers because each antenna views the eye region from a slightly different angle.

The two differential-phase channels use inter-receiver phase differences rather than single-receiver phase. At 60 GHz the radar wavelength is approximately 5 mm, so a head micro-motion of only 1.25 mm along the line of sight shifts the single-receiver phase by π rad. This makes the absolute phase of any individual receiver highly sensitive to involuntary head sway and breathing, and empirically the Pearson correlation ρ between single-receiver phase and horizontal gaze angle is weak ($|\rho| \approx 0.15$). Inter-receiver phase differencing addresses this problem because common-mode path-length changes caused by global head motion, as well as per-receiver hardware phase offsets, cancel in the subtraction. The residual differential signal $\Delta\phi$ encodes small wavefront arrival-angle changes across the receive aperture that are more directly related to the lateral displacement of the scattering point during a gaze shift. Because the sign of this differential path-length change depends on the position of the scattering centre relative to the receive baseline, small variations in head placement between recording sessions can invert the polarity of the $\Delta\phi$ response to a given gaze direction. This session-dependent sign ambiguity motivates a learned classifier that adapts to the prevailing geometry rather than a fixed phase-to-angle rule. It is quantified empirically in Section 4.4. The three receive antennas of the BGT60TR13C are arranged in a short L-shaped aperture with approximately 2.5 mm element spacing. This baseline is too short for classical beamforming at the operating frequency, but it is sufficient for extracting differential phase cues that correlate with gaze direction.

The sixth channel is a differential-amplitude cue formed by subtracting the normalised amplitude of one receiver from another. This inter-receiver amplitude gradient captures spatial energy asymmetries that complement the phase-difference channels. It provides a measure of cross-receiver contrast that does not depend on phase unwrapping or phase-sign ambiguity, and is therefore robust to the phase-related session variability discussed above. Together, the six channels provide a compact representation that remains interpretable at frame resolution after intra-frame chirp averaging. The asymmetry in the feature design—per-receiver values for amplitude, but only inter-receiver differences for phase—reflects the different noise characteristics of the two signal domains at the operating wavelength.

3.4.4 Dataset Construction

After frame-rate feature extraction, supervised samples are formed as fixed-length windows on the radar frame axis. Motion-related intervals used for anchoring are obtained from radar-only detection expressed directly in frame indices. Each event is a contiguous span $[t_{\text{start}}, t_{\text{end}}]$ of frames, produced by mapping the detector's internal time base to the nearest frame boundary and merging adjacent candidates when appropriate. This keeps event definition, feature tensors, and gaze labels on a common discrete timeline.

3.4.5 Radar-event-centred Sampling

For each detected frame interval, let midpoint frame index $t_{\text{c}} = \lfloor \frac{1}{2}(t_{\text{start}} + t_{\text{end}}) \rfloor$ denote the integer centre (or an equivalent tie-breaking rule fixed at dataset build time). A window of frame-window length T consecutive frames centred at t_{c} is extracted, providing symmetric slow-time context around the radar-identified transition at the native frame cadence.

Each feature tensor $\mathbf{X}_i \in \mathbb{R}^{C \times T}$ represents one labelled sample, where channel count $C = 6$ follows Section 3.4. Sessions may differ slightly in usable frame extent near recording boundaries. The frame axis is normalised to a common target length by centre-cropping or edge-padding before concatenation across sessions.

For each radar-centred window, continuous gaze supervision is read from the frame-aligned trajectories at anchor frame t_{c} (or at a fixed offset in frames relative to t_{c} if a post-transition label is preferred). The gaze label vector $\mathbf{y}_i$ for sample i contains

$$\mathbf{y}_i = [\text{azimuth}_i,\ \text{elevation}_i]. \tag{3.9}$$

Direction classes are derived from angular targets through fixed decision rules used during training, while regression heads consume the continuous angles directly.

To balance easy and hard gaze states, each sample receives an importance weight as a monotonic function of angular magnitude. In addition to feature and label tensors, the dataset stores session identity, anchor frame indices, and augmentation offsets, enabling session-wise splits and calibration analyses in later experiments.

3.4.6 Output Format

The final dataset is exported as compressed arrays containing feature windows, angle labels, sample weights, and auxiliary indexing fields. Table 3.3 lists the main exported fields, where sample count N denotes the total number of labelled windows across all sessions. This format supports reproducible training, controlled ablations, and cross-session evaluation without re-running raw signal processing.

Table 3.3: Main fields in the constructed training dataset.

Field	Shape / Type	Purpose
`X`	$(N, T, 6)$ float32	Frame-rate radar feature windows
`y`	$(N, 2)$ float32	Azimuth and elevation targets
`weights`	$(N,)$ float32	Sample importance weighting
`session_idx`	$(N,)$ int16	Session-aware split/evaluation
`frame_center`	$(N,)$ int32	Anchor frame index t_c
`aug_offset`	$(N,)$ int8	Temporal offset augmentation tag

3.5 Model Architecture

3.5.1 Input and Feature Streams

The radar gaze estimator, illustrated in Figure 3.7, is a lightweight dual-stream temporal network that maps an event-centred radar feature window to both a discrete gaze-direction prediction and a continuous gaze-angle estimate. Each input sample is represented as $X_i \in \mathbb{R}^{C_{\text{feat}} \times T}$, where the current dataset uses $T = 50$ frame-level time steps and $C_{\text{feat}} = 6$ channels from the selected range bin. Each frame-level feature is obtained by averaging the selected range-bin response across the 16 chirps within a frame, so T denotes 50 radar frames. The six channels consist of three normalized receiver amplitudes, two filtered inter-receiver phase differences, and one inter-receiver amplitude-difference cue.

The input is separated into an amplitude-like stream and a phase stream before temporal encoding. The amplitude-like stream receives the receiver-amplitude and amplitude-difference cues, while the phase stream receives the phase-difference cues. This separation reflects the different physical roles of the two feature groups: amplitude cues relate to reflected signal strength and energy balance, while phase differences encode relative path-length changes across receivers. Processing them separately allows the first temporal filters to operate on signals with more homogeneous scale and noise characteristics before cross-modal fusion.

3.5.2 Temporal Encoding and Attention Fusion

Each feature stream is encoded by a separate one-dimensional convolutional encoder with the same structure but independent parameters. The encoders use three temporal convolutional layers with kernel sizes 5, 5, and 3, batch normalization, GELU

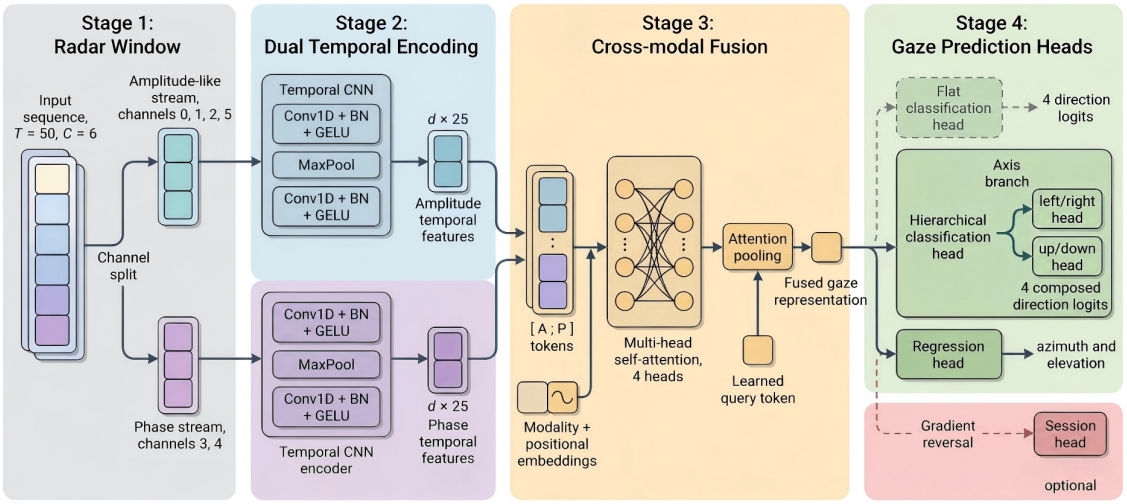


Figure 3.7: Dual-stream hierarchical radar gaze network.

activations, dropout, and one temporal max-pooling operation with stride two. The resulting feature maps can be written as

$$F_i^{(a)} = E_a(X_i^{(a)}), \qquad F_i^{(p)} = E_p(X_i^{(p)}), \qquad F_i^{(m)} \in \mathbb{R}^{d \times T'}, \quad T' = 25,$$

where $X_i^{(a)}$ and $X_i^{(p)}$ are the amplitude-like and phase-difference sub-inputs of sample i, E_a and E_p are the corresponding convolutional encoders, and $F_i^{(m)}$ $(m \in \{a, p\})$ denotes either encoded feature map with channel dimension d and reduced temporal length $T' = T/2 = 25$ after stride-two pooling.

Temporal convolution is used because the radar response around a gaze event is local in time and can be captured with a modest number of parameters [16]. The encoded amplitude-like and phase feature maps are then converted into token sequences. Learned modality embeddings distinguish the two streams, and learned positional embeddings preserve temporal order. A single pre-normalized multi-head self-attention layer with four heads fuses the concatenated token sequence, following the attention mechanism introduced in the Transformer architecture [17]. This fusion step allows the model to represent gaze-dependent patterns that appear jointly across amplitude, phase difference, and receiver-balance cues.

The fused token sequence is compressed by attention pooling. A learned query attends over the token sequence and produces a d-dimensional shared representation f_i. Attention pooling is used instead of simple average pooling because the informative part of an event-centred window is not uniformly distributed over time; the gaze transition and early fixation are expected to carry more information than the pre-event baseline.

3.5.3 Prediction Heads and Model Scale

The shared representation supports both categorical and continuous supervision. The flat classification variant maps f_i directly to four direction logits. The hierarchical variant replaces this with an axis head and two within-axis branch heads, while still returning four composed direction logits. The regression head is shared

by both variants and maps f_i to the two-dimensional angle estimate $[\hat{\theta}_{\mathrm{az},i}, \hat{\theta}_{\mathrm{el},i}]^\top$. Sharing the representation between classification and regression follows the principle of multi-task learning, in which related objectives can benefit from common intermediate features while retaining task-specific output layers [18].
An optional session-discriminator branch can be attached through a gradient-reversal layer for domain-invariance ablations [21]. In the default modality-aware hierarchical configuration with latent feature dimension $d = 32$ and dropout 0.3, the model contains approximately 2.5×10^4 trainable parameters.

3.5.4 Flat and Hierarchical Direction Heads

The model supports two alternatives for the direction-classification head. The default configuration uses a flat four-way classifier that maps the pooled representation f_i directly to logits for $\{\mathrm{LEFT}, \mathrm{RIGHT}, \mathrm{UP}, \mathrm{DOWN}\}$. The hierarchical configuration decomposes the same four-way decision into an axis decision followed by a within-axis branch decision. This decomposition follows the general idea of hierarchical classification, where a structured label space is represented through decisions at multiple levels rather than as a single flat category set [20]. The hierarchy is shallow and physically motivated: gaze stimuli first differ by movement axis, horizontal or vertical, and then by direction within the selected axis.
Let $a_i \in \mathbb{R}^2$ denote the axis logits, where $a_i^{(0)}$ corresponds to the horizontal axis and $a_i^{(1)}$ corresponds to the vertical axis. Let $r_i \in \mathbb{R}^2$ denote the left-right branch logits, and let $u_i \in \mathbb{R}^2$ denote the up-down branch logits. In the fused hierarchical mode, all three heads receive the pooled representation f_i. In the modality-aware hierarchical mode, the axis and up-down heads use an amplitude-stream pooled representation, while the left-right head receives the concatenation of amplitude and phase pooled representations. This option reflects the empirical hypothesis that vertical and horizontal movements may not rely on exactly the same radar cues, while keeping the same class semantics at the output.

3.5.5 Log-Probability Composition

The hierarchical logits are converted back into the four class logits used by the evaluator through log-probability composition. If $\log p_a = \log \mathrm{softmax}(a_i)$, $\log p_r = \log \mathrm{softmax}(r_i)$, and $\log p_u = \log \mathrm{softmax}(u_i)$, the composed class logits are

$$\begin{aligned} \ell_{i,\mathrm{LEFT}} &= \log p_a^{(0)} + \log p_r^{(0)}, \\ \ell_{i,\mathrm{RIGHT}} &= \log p_a^{(0)} + \log p_r^{(1)}, \\ \ell_{i,\mathrm{UP}} &= \log p_a^{(1)} + \log p_u^{(0)}, \\ \ell_{i,\mathrm{DOWN}} &= \log p_a^{(1)} + \log p_u^{(1)}. \end{aligned} \tag{3.10}$$

This composition keeps the external interface identical to the flat classifier. The model still returns four logits in the fixed class order used throughout the dataset and evaluation code. The difference lies in how the logits are parameterized. The flat classifier learns four independent output scores from one shared feature vector, whereas the hierarchical classifier explicitly represents the relation that left and right

share a horizontal parent and up and down share a vertical parent. This structure can be useful when some errors are more naturally axis errors and others are within-axis errors, a distinction that is meaningful in the four-direction gaze protocol.

3.5.6 Hierarchical Training Loss

During training with the hierarchical head, the classification loss is computed as the sum of an axis loss and the branch loss corresponding to the true axis. If $c_i \in \{0, 1, 2, 3\}$ follows the class order $\{\text{LEFT}, \text{RIGHT}, \text{UP}, \text{DOWN}\}$, the axis target is

$$z_i = \mathbb{I}[c_i \geq 2], \tag{3.11}$$

where $z_i = 0$ indicates a horizontal target and $z_i = 1$ indicates a vertical target. Denoting the cross-entropy loss by CE, the per-sample hierarchical classification loss is then

$$\mathcal{L}_{\text{hier},i} = \text{CE}(a_i, z_i) + \mathbb{I}[c_i < 2]\,\text{CE}(r_i, c_i) + \mathbb{I}[c_i \geq 2]\,\text{CE}(u_i, c_i - 2). \tag{3.12}$$

Label smoothing and class weights can be applied to these cross-entropy terms in the same way as in the flat classification setting. The hierarchical loss remains a classification loss over the stimulus-derived class labels, and it does not require camera-angle thresholding. The composed four-way logits are still used for prediction, accuracy, balanced accuracy, and confusion-matrix computation.

3.5.7 Continuous Angle Regression

The regression head is shared by the flat and hierarchical variants. It maps the pooled representation f_i through a two-layer multilayer perceptron and outputs a two-dimensional vector

$$\hat{\mathbf{y}}_i^{\text{reg}} = \begin{bmatrix} \hat{\theta}_{\text{az},i} \\ \hat{\theta}_{\text{el},i} \end{bmatrix}. \tag{3.13}$$

The regression target is the camera-derived angle pair aligned to the radar frame, while the classification target is the discrete stimulus direction. The two heads therefore describe different aspects of the same gaze event, and sharing the fused representation encourages the encoder to retain both categorical and continuous gaze information.

3.6 Training Objective and Optimization

3.6.1 Targets and Losses

The model is trained as a supervised multi-task estimator. For each event-centred radar sample, the input is the radar feature window, the classification target is the stimulus-derived direction label, and the regression target is the camera-derived

azimuth and elevation angle. Radar features are standardized using statistics computed from the training split only, and the regression targets are normalized with the corresponding training-fold mean and standard deviation.
The classification loss is cross-entropy applied to the four direction logits in the flat configuration, or the hierarchical decomposition. Label smoothing is used to reduce over-confident class probabilities [27]. The regression term is the smooth L_1 loss on the normalized azimuth and elevation targets [19], which reduces sensitivity to occasional camera-label noise while retaining a continuous gaze-angle constraint.
The per-sample objective is

$$\mathcal{L}_i = w_i \left(\mathcal{L}_{\mathrm{cls},i} + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg},i} \right), \tag{3.14}$$

where w_i is a clipped and mean-normalized sample weight derived from the magnitude of the camera-estimated gaze displacement, $\mathcal{L}_{\mathrm{cls},i}$ is the flat or hierarchical classification loss, and $\mathcal{L}_{\mathrm{reg},i}$ is the smooth L_1 regression loss. The principal configuration uses $\lambda_{\mathrm{reg}} = 0.2$, so that continuous-angle regression regularizes the shared representation without dominating the direction-classification objective.

3.6.2 Optimization and Model Selection

Optimization is performed with AdamW, which decouples weight decay from the adaptive gradient update [28]. The main configuration uses a learning rate of 10^{-3}, weight decay of 10^{-3}, mini-batches of 32 samples, and a maximum of 80 epochs. A cosine annealing learning-rate schedule is applied over the training horizon [29], and gradients are clipped to a maximum norm of 1.0 before each optimizer step.
Model selection is based on validation balanced accuracy, which gives equal weight to the four gaze directions and is therefore preferable to ordinary accuracy when class frequencies differ across folds. Training stops early if the selected validation score does not improve for 20 consecutive epochs. The selected model is stored together with the training-fold normalization statistics required for later evaluation.

3.6.3 Regularization and Ablation Options

Online augmentation is applied only to training samples. The main configuration uses phase-sign flipping with probability 0.5, which encourages robustness to session-dependent phase polarity. When domain-adversarial training is evaluated, the optional session head adds a session-classification cross-entropy term following Ganin et al. [21]. This branch is treated as a domain-generalization ablation and is not part of the default objective.

3.7 Evaluation Protocol and Inference Pipeline

3.7.1 Deployment Setting

The final deployment path of the system is radar-only. In this setting, the camera and the stimulus protocol are used during dataset construction, supervision, and evaluation, but they are not used as input modalities when a trained model is

applied to a new recording. The inference pipeline has two functional stages, as illustrated in Figure 3.8. First, candidate eye-movement intervals are detected from the radar signal itself. Second, each candidate interval is converted into the same fixed-length feature-window representation used during supervised learning and is passed to the trained gaze estimator.

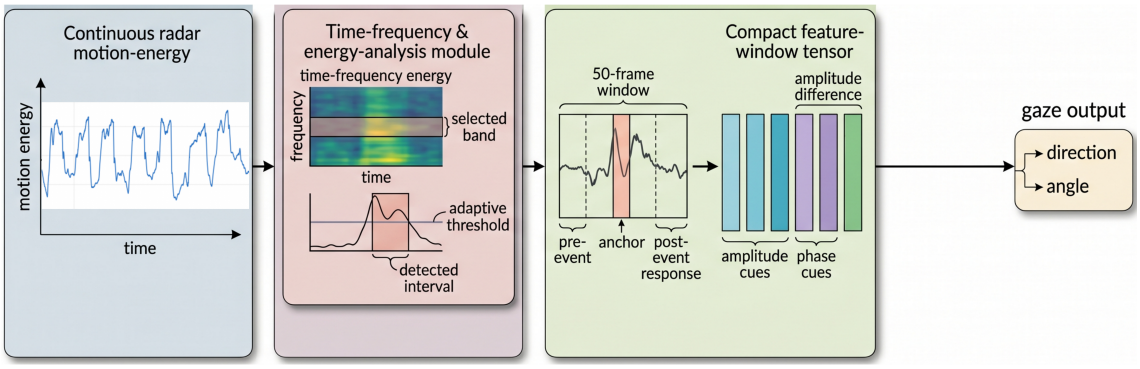


Figure 3.8: Conceptual radar-only inference pipeline.

3.7.2 Radar Event Proposal

The radar-only stage begins by treating fast eye movements as short-duration changes in the radar return. For event proposal, the selected range-bin response can be analysed as a chirp-level slow-time sequence, preserving the higher temporal sampling available within each radar frame. These changes are more visible after transforming this signal into a local time-frequency representation and reducing the response within the frequency band associated with rapid eye motion. Let $S(\tau, f)$ denote the time-frequency representation of the selected radar return at time τ and frequency f, and let $\mathcal{B}$ denote the analysis band used for rapid motion. The band-limited motion energy $E(\tau)$ and the corresponding adaptive event set are expressed as

$$E(\tau) = \sum_{f \in \mathcal{B}} |S(\tau, f)|^2, \qquad \mathcal{I} = \{\tau \mid E(\tau) > \mu_\tau + \kappa\sigma_\tau\}. \tag{3.15}$$

In Equation 3.15, μ_τ and σ_τ are local background statistics of the energy curve, and κ controls the sensitivity of the adaptive threshold. Consecutive samples in $\mathcal{I}$ form candidate eye-movement intervals after duration filtering and temporal merging. The purpose of this stage is not to determine gaze direction, but to replace camera or stimulus timing with candidate event timing inferred from radar alone.

3.7.3 Window Construction and Prediction

Once an interval has been detected, it is resolved to an anchor frame on the radar time axis. A fixed-length window is then constructed around this anchor, with a short pre-event segment and a post-event response segment. The feature construction follows the same convention as the training data. The gaze estimator receives chirp-averaged frame-level features containing normalized receiver amplitudes, inter-receiver phase differences, and an amplitude-difference cue. Baseline correction and

phase filtering are applied so that the model receives relative motion cues rather than absolute offsets. This consistency between training and inference is important because the neural estimator is trained on event-centred windows rather than on arbitrary segments of the continuous radar stream.

The trained estimator is applied only after the radar interval has been converted into the expected feature representation. The feature window is normalized using the statistics saved during training, and the angle output is converted back from normalized target space to physical azimuth and elevation units using the stored target statistics. Session-level normalization statistics are estimated from radar-only data during deployment.

3.7.4 Interval-Level Output

The output for each valid radar interval consists of a probability distribution over the four gaze directions and a two-dimensional angle estimate. The current design should therefore be interpreted as interval-level radar gaze inference rather than dense frame-by-frame gaze tracking. It first detects candidate eye-movement intervals from radar and then estimates gaze direction and angle for those intervals. This separation keeps the event-proposal problem explicit while allowing the trained estimator to focus on the radar response around each detected eye movement.

3.7.5 Evaluation Settings

The evaluation protocol is organized around the distinction between within-session learnability, cross-session generalization, and calibration-assisted deployment. This separation is necessary because radar gaze estimation is strongly affected by session-specific factors such as subject posture, sensor placement, range response, and phase behaviour. A single random split would therefore overstate the generalization capability of the system by mixing samples from the same recording condition across training and validation. The protocol follows the general motivation of cross-validation for model assessment, where the evaluation split should reflect the intended use condition of the model [30].

Table 3.4 summarizes the evaluation settings used in this thesis. The event-level split groups samples by gaze event so that windows derived from the same event are not divided between training and validation. Since different events from the same session may still appear on both sides of the split, this setting mainly verifies that the radar representation and labels are mutually informative. The leave-one-session-out protocol is the principal generalization test. In each fold, one entire recording session is held out, and the model is trained on the remaining sessions. The calibration-assisted setting is reported separately. In this setting, a small number of labelled target-session events are added before evaluation, which quantifies the benefit of a short session-specific calibration step rather than measuring strict session-independent performance.

Table 3.4: Evaluation protocols used for radar gaze estimation.

Protocol	Purpose	Interpretation
Event-level split	Tests whether event-centred radar windows contain learnable gaze information within the same recording distribution.	Optimistic upper bound, not evidence of cross-session deployment performance.
Leave-one-session-out	Holds out one full recording session and trains on the remaining sessions.	Main offline test of cross-session generalization.
Calibration-assisted split	Adds a small class-balanced subset from the held-out session before evaluating the remaining target-session events.	Calibrated setting, not pure leave-one-session-out generalization.
Feature and model ablations	Repeats the same split and metric definitions while changing feature sets or model components.	Design comparison under controlled validation conditions.

3.7.6 Classification Metrics

Classification performance is reported using ordinary accuracy, balanced accuracy, and the confusion matrix. Balanced accuracy is used as the primary scalar classification score because it gives equal weight to all four gaze directions and is less sensitive to class-frequency variation [31]. If confusion matrix M has class count $C_{\text{cls}} = 4$, balanced accuracy BAcc is defined as

$$\text{BAcc} = \frac{1}{C_{\text{cls}}} \sum_{k=1}^{C_{\text{cls}}} \frac{M_{kk}}{\sum_{j=1}^{C_{\text{cls}}} M_{kj}} \tag{3.16}$$

The confusion matrix complements Equation 3.16 by showing whether errors mainly occur between axes or within the same axis. This matters for interpreting the hierarchical direction head, because a horizontal-versus-vertical error has a different meaning from a left-versus-right or up-versus-down error even when the scalar accuracy is unchanged.

3.7.7 Angle Metrics and Reporting

Continuous gaze-angle performance is evaluated after the regression outputs are converted back to physical angular units. The azimuth and elevation errors are reported separately in degrees, because the two axes may be affected differently by camera-label noise, radar geometry, and session-dependent phase behaviour. For sample count N, let $\hat{\theta}_{\text{az},i}$ and $\hat{\theta}_{\text{el},i}$ denote the predicted azimuth and elevation angles for sample index i, and let $\theta_{\text{az},i}$ and $\theta_{\text{el},i}$ denote the corresponding reference angles. The mean absolute errors are

$$\mathrm{MAE}_{\mathrm{az}} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_{\mathrm{az},i} - \theta_{\mathrm{az},i}\right|, \qquad \mathrm{MAE}_{\mathrm{el}} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_{\mathrm{el},i} - \theta_{\mathrm{el},i}\right|. \tag{3.17}$$

The classification labels are derived from the controlled stimulus direction, whereas the continuous angle targets are derived from the camera-based gaze estimate. Reporting both sets of metrics therefore reflects the two forms of supervision used in the system. In Chapter 4, event-level results, strict leave-one-session-out results, calibration-assisted results, and ablation results should be presented separately so that within-session learnability, cross-session robustness, and calibration benefit are not conflated.

3.7.8 Compared Methods

The proposed DualStreamGazeNet is compared against two tiers of baseline models to isolate the contributions of temporal modelling and modality-aware design. All baselines are evaluated under the same leave-one-session-out protocol with identical folds and label source.

The non-temporal tier applies classical classifiers to temporally pooled features. Each event window of $T = 50$ frames and $C = 6$ channels is reduced to a 36-dimensional vector containing the mean, standard deviation, minimum, maximum, peak-to-peak range, and median of each channel. Three classifiers operate on this representation. The first is a support vector machine with an RBF kernel (SVM-RBF), using regularisation parameter $C = 10$ and kernel-width parameter $\gamma =$ `scale`. The second is a random forest with 500 trees and `min_samples_leaf` $= 2$. The third is a two-layer multilayer perceptron with hidden layers of width 128 and 64, trained with early stopping. All three use standard scaling and class-balanced weighting. By discarding temporal ordering, this tier tests whether pooled summary statistics carry sufficient gaze information for cross-session classification.

The temporal tier is a single-stream convolutional network (SingleStreamCNN) that processes the full $T \times C = 50 \times 6$ input through three one-dimensional convolutional blocks (batch normalisation, GELU activation, max-pooling), followed by global average pooling and a flat four-class classification head with an auxiliary angle-regression output. It uses the same training procedure as DualStreamGazeNet (per-fold normalisation, AdamW optimisation, cosine learning-rate annealing, and early stopping) but processes all six radar channels jointly from the first layer onward, without amplitude-phase separation, cross-modal attention, or the modality-aware hierarchical head. This tier isolates the effect of temporal modelling from the effect of modality-aware architecture.

4

Results

This chapter reports the empirical evaluation of the proposed radar-based gaze estimation pipeline. We first summarise the acquisition conditions and dataset construction; subsequent sections present ground-truth diagnostics, signal-level observations, model performance, and ablations.

4.1 Experimental Setup

4.1.1 Participants and Recording Campaign

All radar–camera recordings used in this thesis were collected by the two co-authors in a shared laboratory protocol. In total, **15 sessions** were acquired under consistent hardware and software settings, yielding a compact but fully traceable corpus for supervised learning and offline evaluation. Each session comprised approximately 3–4 minutes of active stimulus logging, during which the participant executed the full trial sequence of centre-hold, outward saccade, peripheral fixation, and return. Because both authors alternated as operator and participant, the dataset reflects a two-person gaze and physiology distribution; findings should therefore be read as a proof-of-concept study rather than a population study.

4.1.2 Acquisition Hardware and Geometry

Millimetre-wave sensing was performed with an Infineon BGT60TR13C FMCW radar front-end operated through the vendor acquisition stack. Colour video for gaze supervision was captured with an Intel RealSense RGB-D camera at 640×480 resolution and nominally 30 fps. The radar was programmed for 50 Hz frame updates with 16 chirps per frame and 400 fast-time samples per chirp at an IF sampling rate of 500 kHz, with a swept bandwidth consistent with the nominal 60 GHz-band FMCW configuration used during collection.

Participants were seated at desktop range in front of a 1920×1080 display with the chin stabilised on an adjustable rest to limit voluntary head translation. The radar module was rigidly mounted with its boresight directed toward the facial region; exact stand-off and tilt are documented in the session metadata and were kept as consistent as practical across sessions.

4.1.3 Visual Tasks and Logging

Each acquired session stored the radar tensor, time-stamped camera frames with online gaze estimates, and structured logs for stimulus segmentation. These modalities provide complementary supervision: dense gaze angles for regression, and protocol metadata for qualitative alignment and optional stratified analysis.

4.1.4 Software Environment

Recordings were controlled from a laboratory workstation under Windows. Radar and camera streams were acquired concurrently; host timestamps bracket hardware reads to support offline temporal association between modalities. No real-time closed-loop control was required for the experiments reported here.

4.1.5 Ethical and Data-handling Note

The corpus contains biometrically informative video and radar traces from the two student authors only. Data are stored locally under access control; any future release beyond the thesis context would require renewed consent and anonymisation policies.

4.2 Dataset Statistics

After session discovery and quality checks, 15 sessions entered the dataset construction pipeline described in Chapter 3. The final supervised set is obtained by converting each accepted session into fixed-length radar feature windows with aligned gaze angles. Table 4.1 lists aggregate quantities.

Table 4.1: High-level dataset statistics

Quantity	Value
Number of sessions	15
Number of supervised samples N	1594
Temporal window length T	50
Feature channels C	6
Radar frame rate	50 Hz
Camera frame rate	30 fps

The 1,594 supervised windows correspond to roughly 400 samples per cardinal class (Table 4.2), which is consistent with small-sample radar micromotion studies at comparable scales. Each window is centred on an independently detected gaze event and draws from a 1-second feature epoch. For a proof-of-concept study restricted to two participants, this count is sufficient to train a compact attention model and evaluate cross-session generalisation, but is insufficient to make population-level claims.

Continuous supervision consists of azimuth and elevation pairs expressed in radians after temporal alignment to the radar timeline. Cardinal-direction classes, when

used, are derived from these angles under fixed angular thresholds in the training objective, or from stored auxiliary fields when available. Table 4.2 should report class counts and the fraction of samples labelled as "other" or excluded after filtering.

Table 4.2: Cardinal-class distribution after dataset construction

Class	Count	Proportion (%)
Left	367	23.0
Right	381	23.9
Up	432	27.1
Down	392	24.6
Other / excluded	22	1.4

4.3 Camera Ground Truth Quality

The camera-derived gaze estimates were evaluated before they were used as continuous supervision for the radar model.

4.3.1 Aligned Label Availability

The first requirement is that camera-derived labels remain available after synchronization to the radar time base. Across the 15 sessions used for the event-based dataset, 205,902 of 205,904 radar-aligned frames had finite azimuth and elevation estimates, corresponding to 99.999% valid aligned labels. Rounded to one decimal place, the mean frame-level validity is therefore 100.0%, as shown in Figure 4.1. The two missing aligned frames occurred in a single session and did not affect the selected event windows.

This result narrows the relevant source of label uncertainty. Missing landmarks and timestamp mismatch are not the dominant limitations in the retained recordings, so later radar errors cannot be explained simply by absent camera labels. At the same time, availability is only a completeness measure. A valid frame can still contain angular bias or short-term jitter, so additional checks are needed before treating the camera stream as a useful regression reference.

4.3.2 Directional and Event-Level Consistency

The second requirement is that the camera-derived angles follow the directional structure of the stimulus protocol. Figure 4.1 shows that left and right stimuli separate primarily along azimuth, with median azimuth values of approximately $-30.7°$ and $26.3°$. Upward and downward stimuli separate primarily along elevation, with median values of approximately $-8.4°$ and $12.8°$, following the sign convention in Section 3.3. This separation supports the use of camera-derived angles as continuous regression targets because the main variation occurs along the expected gaze axis.

The same figure also shows that the separation is not perfectly axis-pure. Cross-axis components remain visible, and the vertical angular range is smaller than the

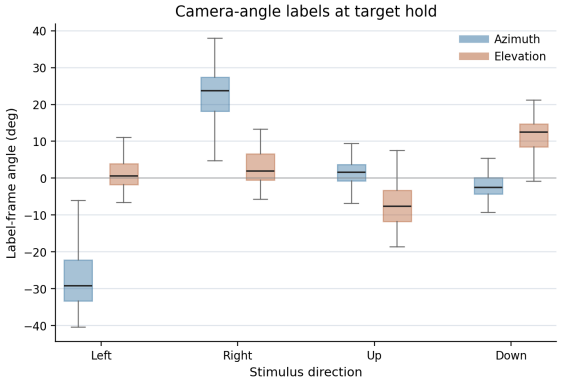


Figure 4.1: Session-level camera ground-truth quality summary. It shows the camera-derived angles at the selected label frames, grouped by stimulus direction.

horizontal range. These effects are consistent with a monocular landmark-based reference collected without full eye-tracker calibration. They justify two methodological choices used later in the thesis: the discrete direction label is taken from the known stimulus direction rather than from thresholded camera angles, and azimuth and elevation errors are reported separately rather than collapsed into a single scalar.

An event-level quality screen retained possible directional gaze events after requiring a clear camera-observed transition and a settled label frame. The rejected events were primarily cases in which the camera trajectory did not show a sufficiently distinct saccadic movement, not cases with missing camera measurements. This distinction is important for interpreting the ablation study in Section 4.9. Filtering increases apparent camera-motion clarity, but it also changes the training distribution by removing events that may still contain valid stimulus-driven radar responses.

4.3.3 Screen-Target Validation

Geometric plausibility was further evaluated for one participant using known two-dimensional screen targets. The camera representation was mapped to normalized screen coordinates with a participant-specific calibration, and target-wise generalization was evaluated by withholding one target location at a time. For a known target (x_i, y_i) and a predicted target $(\hat{x}_i, \hat{y}_i)$, the screen error was expressed as a visual angle,

$$e_i = \frac{180}{\pi} \arctan\left(\frac{\sqrt{[(\hat{x}_i - x_i)W]^2 + [(\hat{y}_i - y_i)H]^2}}{D}\right), \tag{4.1}$$

where W and H denote the physical screen width and height, and D denotes the viewing distance. This metric expresses screen-coordinate error in degrees of visual angle, making it comparable to the angular labels used in the radar experiments.

Figure 4.2 shows that the camera reference preserves the spatial ordering of the screen targets, with an in-sample mean angular error of $2.51° \pm 1.33°$ and a leave-one-point-out error of $3.05° \pm 1.55°$. The moderate increase under target-wise cross-

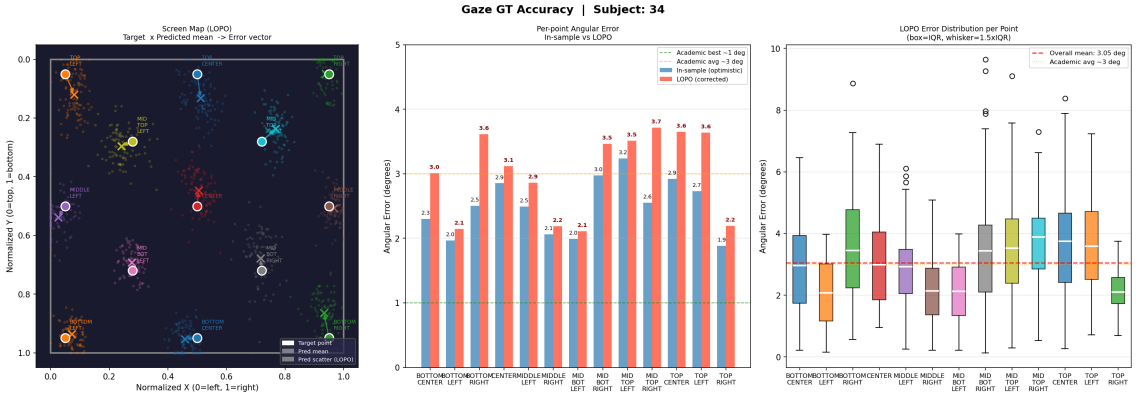


Figure 4.2: Screen-target validation of the camera-derived ground truth for one participant. The left panel compares known target positions with leave-one-point-out predicted means and frame-level predictions. The centre panel compares in-sample and leave-one-point-out angular errors across target positions. The right panel shows the leave-one-point-out error distribution for each target position.

validation indicates that the mapping is strongest when interpolating within the observed calibration geometry and less reliable when extrapolating to a withheld location. Several peripheral or upper-screen targets show broader residual distributions, which is consistent with the limited geometric constraint of monocular landmark-based gaze estimation.

4.3.4 Implications for Radar Evaluation

Taken together, the camera reference is adequate for its role. The direction-wise angle distributions show that the reference preserves the intended stimulus geometry, which supports its use for continuous regression. The screen-target validation further suggests that errors are on the order of a few degrees under controlled conditions, but with location-dependent bias.

These findings motivate the label design used in the radar experiments. Stimulus-derived labels are used for direction classification because they avoid introducing camera-threshold noise into the class target. Camera-derived angles are used for regression because they preserve continuous variation that is not available from the stimulus direction alone.

4.4 Radar Signal Characteristics

Before reporting quantitative model performance, we summarise qualitative properties of the recorded waveforms and derived features. The goal is to show (at the level of range-resolved energy, slow-time amplitude and phase-derived displacement, spectro-temporal structure, and complex baseband geometry) that radar observables co-vary with gaze kinematics in ways consistent with micromotion-dominated scattering at desktop range. The figures in this section were produced with the same offline visualisation pipeline used during acquisition review, and are intended

as interpretability evidence rather than evaluation metrics.

4.4.1 Range-resolved Energy and Bin Selection

Figure 4.3 summarises, for one representative recording, how temporal variability of the magnitude response is distributed across apparent range after standard fast-time spectral estimation and per-chirp stacking. The variance profile is strongly peaked at a short stand-off distance beyond the immediate near-field region. Most range cells remain quiescent, while a narrow band exhibits elevated slow-time dynamics. This pattern motivates a session-adaptive choice of analysis range gate. The operating bin is selected as the range index that maximises this session-level variability criterion, so that subsequent slow-time and time–frequency analyses focus on the distance interval where micromotion rather than static leakage dominates.

Figure 4.4 complements the variance view with a single-frame range–amplitude profile at the same nominal geometry. Strong energy adjacent to zero range reflects fixed clutter and sensor-proximate scattering, while the automatically selected gate (marked for reference) lies on the flank of that structure, where the echo is still energetic yet more sensitive to small mechanical perturbations. Mid-range sidelobes and a gently sloping noise floor show that the desktop scene is weakly structured. Together, a peaked variance curve and a stable in-range profile support treating the selected cell as a reliable observation point for eye-related micromotion throughout the session.

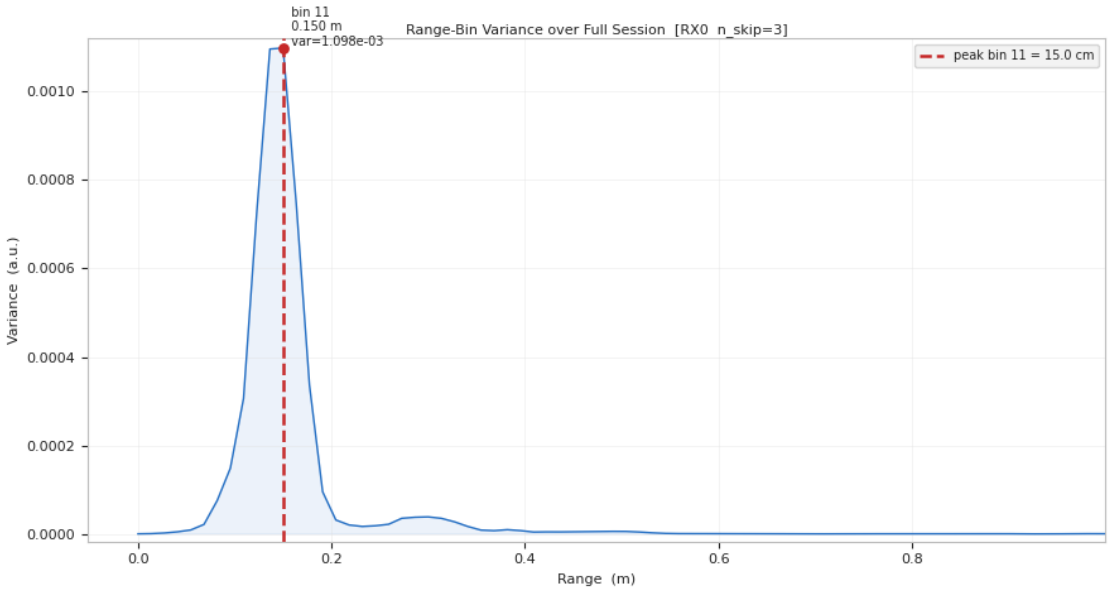


Figure 4.3: Session-level variability of the magnitude response across range cells. A dominant peak indicates the distance interval where slow-time modulation is strongest, guiding automatic selection of the analysis bin.

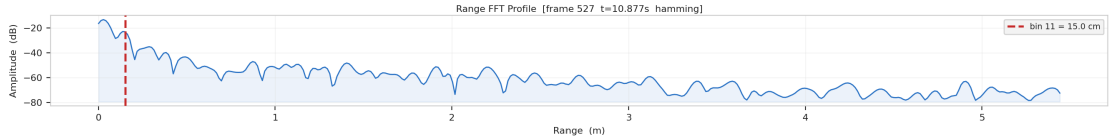


Figure 4.4: Illustrative range–amplitude profile within one frame, showing near-field clutter, secondary reflectors, and the session-adaptive range gate used for downstream slow-time inspection.

4.4.2 Slow-time Amplitude and Phase Co-travel with Gaze

Figure 4.5 shows the gaze angles in the horizontal and vertical directions alongside the radar-derived slow-time trajectories in the selected range cell. The top row shows phase unfolded and scaled to a displacement proxy, and the bottom row shows amplitude. During quasi-fixation episodes, amplitude drifts slowly while the displacement trace remains largely oscillatory but bounded. Around rapid gaze steps, both channels show coordinated transients whose onsets and decays follow the steep segments of the gaze trajectory, to within the alignment uncertainty of the multimodal setup. These overlays should not be read as instantaneous causal statements at the single-frame level. Rather, they establish temporal coherence between micromotion-sensitive radar observables and oculomotor events at the scale of tens of milliseconds, the regime directly relevant to event-centred windowing in Chapter 3.

Additional sessions are documented in Appendix A.1 with the same layout.

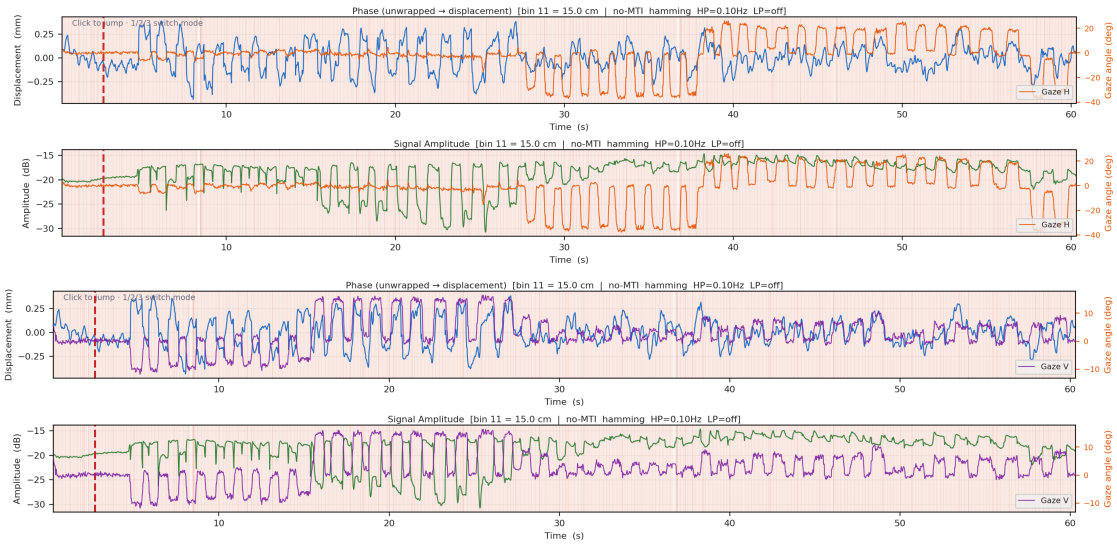


Figure 4.5: Radar slow-time displacement (top) and amplitude (bottom) overlaid with horizontal (upper pair) and vertical (lower pair) gaze angles on a common timeline for one representative session.

4.4.3 Directional Structure in the Complex IQ Plane

Figure 4.6 aggregates short excerpts of the complex baseband response at the analysis range cell, plotted in the in-phase versus quadrature (IQ) plane for stimulus-defined horizontal and vertical targets. Within each excerpt, samples are coloured by a coarse segmentation of the camera-based gaze trace. The three phases are outward motion from central fixation toward the peripheral target (blue), a quasi-static hold at the target (green), and the return movement toward centre (orange). Outward and return phases trace extended arcs or loops in IQ space, whereas hold intervals form tighter clusters, consistent with reduced micromotion during fixation. Qualitatively, the four directions are not identical in IQ geometry. Vertical and horizontal protocols excite distinguishable trajectory shapes across repeated trials, which supports using multi-channel radar features that retain both magnitude and phase information rather than reducing the echo to a single scalar per time step.

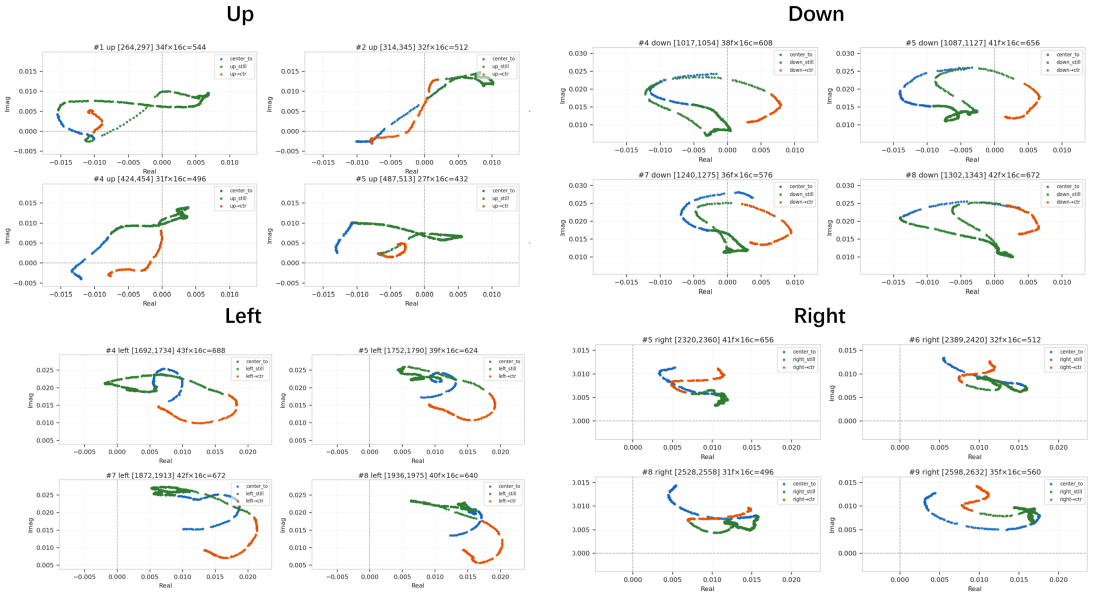


Figure 4.6: Representative complex-baseband trajectories in the IQ plane for four cardinal gaze directions. Point colours denote camera-segmented phases: outward saccade from centre to target (blue), stable hold at the target (green), and return saccade toward centre (orange).

4.4.4 Micro-Doppler and Spectro-temporal Structure

Figure 4.7 shows the short-time spectrum of the complex slow-time series at the analysis range cell, with gaze angles superimposed on the same horizontal axis. Residual low-frequency ridge energy reflects slowly evolving clutter and bulk motion, whereas brief vertical "flares" of broadband energy coincide with rapid gaze transitions and largely recede during fixation intervals. This spectro-temporal behaviour is consistent with the radar-only event detector's underlying assumptions. Band-limited energy aggregation in the frequency–time plane is sensitive to saccade-scale bursts while remaining comparatively quiescent during steady fixation, supporting the use

of radar-detected intervals as anchors for supervised samples as described in Section 3.4.4.

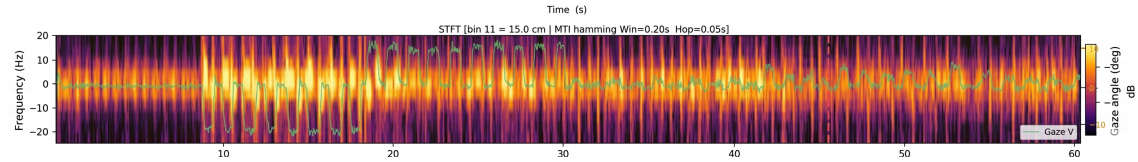


Figure 4.7: Time–frequency energy at the selected range cell with overlaid gaze, illustrating micro-Doppler-like bursts timed with rapid gaze changes.

4.4.5 Direction-dependent Feature Behaviour in Event Windows

The preceding subsections examine radar observables at the session level. Figure 4.8 complements that view by showing the extracted model-input features within event-centred windows, aggregated by stimulus direction for a single recording session. The $\Delta\phi_{10}$ channel (b), whose receive pair RX1–RX0 spans a diagonal baseline with both horizontal and vertical components, exhibits the clearest directional structure: upward and downward gaze produce opposite-sign excursions during the event, while horizontal directions remain closer to zero. The amplitude channels (a) show that all four directions produce a visible signal change, with vertical gaze events reaching somewhat larger peak amplitudes. The $\Delta\phi_{20}$ channel (c), corresponding to the purely horizontal RX2–RX0 baseline, carries less directional contrast in this session, consistent with horizontal gaze discrimination being more sensitive to session-specific sensor-to-face geometry. These patterns motivate the dual-stream architecture and modality-aware classification head described in Chapterr 3 .

To move beyond the qualitative patterns visible in Figure 4.8, the directional contrast of the $\Delta\phi_{10}$ channel is measured quantitatively (Figure 4.9). For each event the signed excursion is defined as the mean $\Delta\phi_{10}$ in the event region (frames 15–35) minus the pre-event baseline (frames 0–9). Cohen's d between leftward and rightward events computed within individual recording sessions yields a median $|d| = 0.55$ across the fifteen sessions (range 0.11–0.94), indicating a medium-to-large effect. Although the RX1–RX0 pair spans a diagonal baseline and shows a comparable up–down contrast (median $|d| = 0.52$), it provides the stronger leftward–rightward separation of the two phase-difference channels: the purely horizontal RX2–RX0 pair reaches only $|d| = 0.36$, which motivates the use of $\Delta\phi_{10}$ here. When all sessions are pooled, however, $|d|$ falls to 0.18: the polarity of the phase-difference response reverses in approximately half of the sessions, so opposite-sign effects partially cancel.

The within-session Pearson correlation between the excursion and continuous azimuth angle has a median $|r| = 0.13$, with nine of fifteen sessions exceeding $|r| = 0.10$ and a maximum of 0.32. This pattern is consistent with the physical expectation that the differential path length between RX1 and RX0 depends on the scatterer position relative to the receive baseline: re-seating the participant introduces small geometric changes that alter the sign mapping. The session-dependent polarity confirms that

the inter-receiver phase difference carries substantial directional information that a session-aware learned model can exploit, while a pooled linear statistic understates the feature's discriminative value.

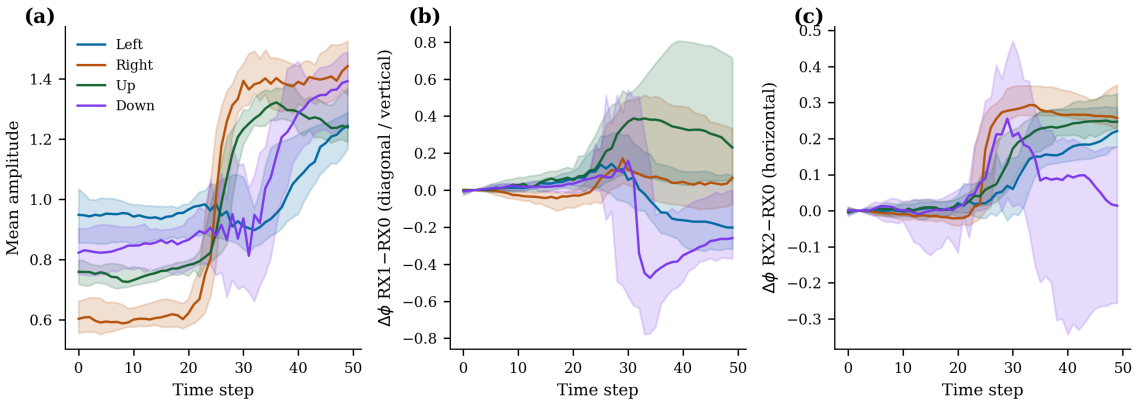


Figure 4.8: Median radar feature traces within event-centred windows for the four stimulus directions, computed from a single recording session. Shaded bands indicate the interquartile range across events. (a) Mean normalised amplitude across the three receivers. (b) Inter-receiver phase difference (RX1−RX0, diagonal baseline). (c) Inter-receiver phase difference (RX2−RX0, horizontal baseline).

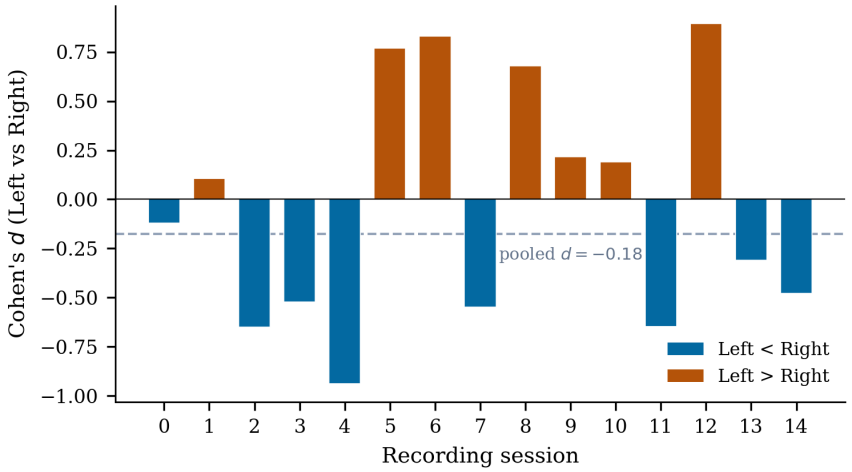


Figure 4.9: Per-session Cohen's d for the $\Delta\phi_{10}$ (RX1−RX0) excursion between leftward and rightward gaze events.

4.5 Radar-only Event Detection

The classification, regression, calibration, and ablation results, which will be presented in following sections, are all reported on event-centred radar windows whose anchors are obtained from the radar-only event proposal stage of Section 3.7. This section quantifies how reliably that stage recovers eye-movement events when compared against the camera- and stimulus-derived ground-truth (GT) events defined in Section 4.3, completing the empirical evaluation of the radar-only inference path.

4.5.1 Evaluation Protocol

For each session, the radar-only detector of Equation 3.15 is applied to the continuous radar recording, producing a set of candidate event onsets $\{\hat{t}_k\}$. The reference set $\{t_\ell\}$ is constructed from the camera-derived gaze trajectory around each commanded stimulus direction using the same onset-fraction rules used during dataset construction. A radar-proposed onset $\hat{t}_k$ is matched to a reference onset t_ℓ when $|\hat{t}_k - t_\ell| \leq \tau_{\text{match}}$, with $\tau_{\text{match}} = 500\,\text{ms}$. Matched pairs are counted as true positives, unmatched predictions as false positives, and unmatched reference events as false negatives. Per-session precision, recall, and F_1 are computed from these counts, together with the mean absolute onset-time error on matched pairs. Each reference event is matched to at most one predicted onset (nearest-first assignment), so the metric is not inflated by duplicate detections.

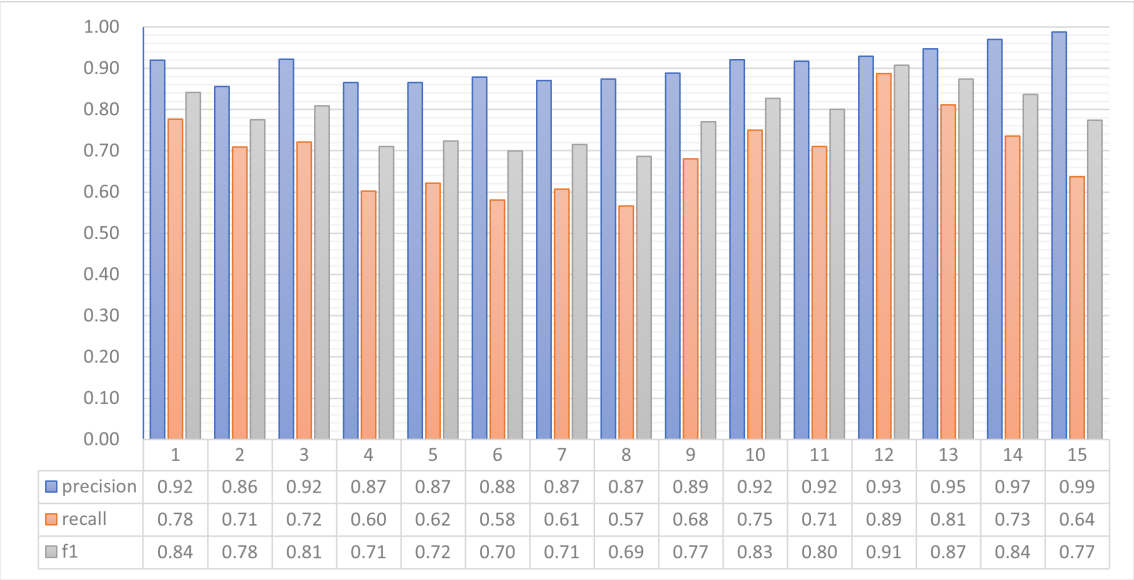
4.5.2 Per-session and Aggregate Performance

Figure 4.10 summarises the per-session results on the 15 evaluation sessions. The detector reaches a mean precision of 0.907 with a tight per-session spread (0.86–0.99), indicating that the vast majority of radar-proposed events correspond to genuine gaze transitions and that the operating point is comparable across sessions without per-session retuning. Recall is lower and more session-dependent, with a mean of 0.693 and a range of 0.57–0.89, showing that some reference events go undetected at the chosen threshold. The mean absolute onset-time error on matched pairs is 79.8 ms, well below the 1.00 s training window length ($T = 50$ frames at the 50 Hz radar frame rate) used by the gaze estimator. The aggregate F_1 across the 15 sessions is 0.783.

4.5.3 Interpretation

These numbers should be read jointly with the downstream classification results in Section 4.3. The high and stable precision means that the windows fed to the gaze estimator are predominantly anchored on real eye-movement events rather than on head-motion artefacts or sensor drift; this is consistent with the strong event-level classification performance, which would not be achievable on systematically mis-anchored windows. The 79.8 ms mean onset error is small relative to both the 1.00 s analysis window and the typical post-saccadic settling duration, so the residual detection jitter does not push the informative portion of the event outside the model's input.

Recall variability has a more direct deployment cost. The four lowest-recall sessions miss roughly 40% of the reference events at the present operating point, which would manifest as missed glances in a continuous driver-monitoring scenario. The high-precision and moderate-recall regime observed here is the expected behaviour of a conservative threshold; the operating point of Equation 3.15 can be shifted toward higher recall at the cost of admitting more false alarms whenever the downstream application can tolerate them.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
precision	0.92	0.86	0.92	0.87	0.87	0.88	0.87	0.87	0.89	0.92	0.92	0.93	0.95	0.97	0.99
recall	0.78	0.71	0.72	0.60	0.62	0.58	0.61	0.57	0.68	0.75	0.71	0.89	0.81	0.73	0.64
f1	0.84	0.78	0.81	0.71	0.72	0.70	0.71	0.69	0.77	0.83	0.80	0.91	0.87	0.84	0.77

Figure 4.10: Per-session precision, recall, and F_1 of the radar-only event proposal stage evaluated against camera- and stimulus-derived ground-truth events on the 15 recording sessions. A predicted onset is counted as a true positive when it lies within $\tau_{\text{match}} = 500\,\text{ms}$ of a reference onset. Mean values across sessions are $P = 0.907$, $R = 0.693$, $F_1 = 0.783$, with a mean absolute onset-time error of 79.8 ms on matched pairs.

4.6 Gaze Estimation Results

This section evaluates whether the proposed model can infer gaze from radar-derived feature windows alone. Camera-derived angles and stimulus directions are used only as evaluation references, not as model inputs. The main results are summarized in Table 4.3, and the interpretation below focuses on the comparisons between evaluation protocols rather than on isolated metric values.

Table 4.3: Gaze estimation performance under the main evaluation protocols. Balanced accuracy is the mean per-class recall over the four gaze directions. Macro F1 is the unweighted mean of per-class F1-scores. Azimuth and elevation errors are mean absolute errors in degrees.

Evaluation protocol	Balanced acc.	Std	Macro F1	Az MAE	El MAE
Event-level split	85.7%	±1.1 pp	85.6%	7.82°	3.15°
Leave-one-session-out	71.5%	±14.0 pp	71.5%	11.23°	4.22°
Calibration-assisted, $K = 1$	72.2%	±12.2 pp	72.1%	11.45°	4.30°
Calibration-assisted, $K = 3$	78.2%	±10.9 pp	78.1%	10.13°	3.78°
Calibration-assisted, $K = 5$	82.1%	±11.7 pp	82.0%	9.58°	3.52°

4.6.1 Feasibility and Cross-Session Gap

The event-level split establishes that the radar representation contains gaze discriminative information under matched recording conditions. Balanced accuracy reaches 85.7%, and the standard deviation across folds is only 1.1 percentage points. This low variance is important because it indicates that the result is not driven by a favourable train-validation partition. It also supports the basic feasibility of the event-centred representation: when the session distribution is shared, the model can separate the four gaze directions consistently.
The leave-one-session-out result changes the interpretation. Balanced accuracy decreases to 71.5%, a 14.2 percentage-point reduction that is far larger than the event-level fold variation. This comparison shows that the main performance bottleneck is cross-session shift rather than a failure to learn gaze-related radar patterns. The same conclusion appears in the regression metrics: azimuth error increases by 3.41°, while elevation error increases by 1.07°. The larger horizontal degradation suggests that session-specific radar geometry affects horizontal gaze estimation more strongly than vertical gaze estimation, a pattern examined further in the calibration and ablation analyses.

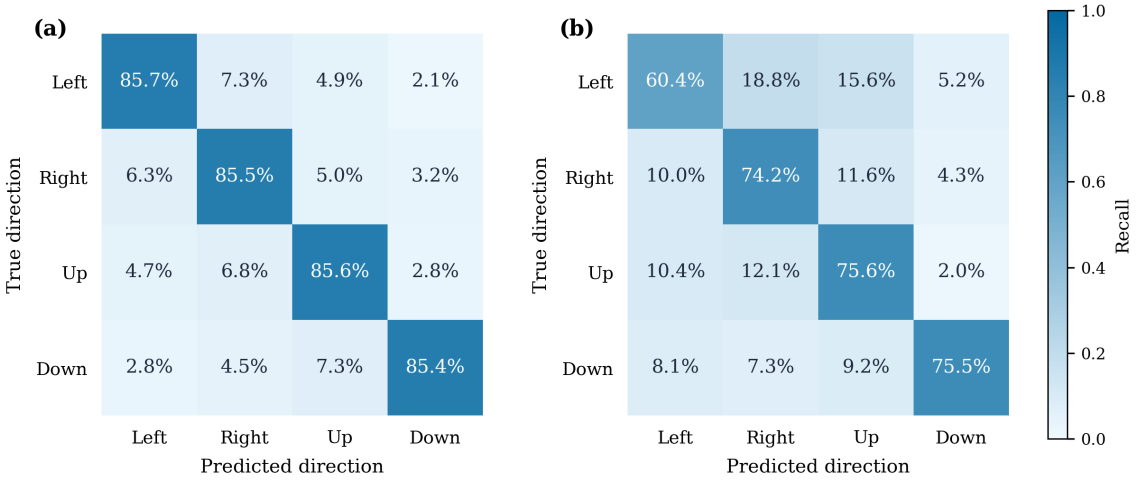


Figure 4.11: Aggregated confusion matrices for the event-level split and the leave-one-session-out protocol. Cell values report row-normalised recall and sample counts.

4.6.2 Direction-Wise Error Structure

Figure 4.11 shows that the cross-session gap is not distributed evenly across directions. In the event-level split, the four classes have similar recall, which suggests that the dataset does not contain an intrinsically unlearnable gaze direction under matched conditions. In the leave-one-session-out setting, however, leftward gaze becomes the weakest class, with recall falling to 60.4%, while rightward, upward, and downward gaze remain around the mid-70 percent range.
This asymmetry is more informative than the aggregate accuracy drop. If all directions degraded similarly, the main explanation would be a general loss of signal

quality in unseen sessions. Instead, the error distribution indicates that some gaze motions are less stable across session geometries. Leftward gaze produces both within-axis confusion and cross-axis confusion, which means the model sometimes loses the horizontal polarity and sometimes fails to preserve the movement axis itself. Chapter 5 discusses the possible geometric causes, but the direct result here is that robust radar gaze estimation cannot be judged only by mean accuracy. Direction-specific recall is necessary for identifying which gaze movements are most sensitive to session shift.

Table 4.4 complements the recall-focused confusion matrices by reporting per-class precision and F1-score across the three main evaluation settings. Under the event-level split, precision and recall are closely matched for all four directions, producing per-class F1-scores in the 83.8% to 88.2% range. This balance indicates that no single direction is systematically over- or under-predicted when the session distribution is shared between training and evaluation.

Under the leave-one-session-out protocol, the per-class metrics diverge. Leftward gaze exhibits the largest recall–precision gap among the horizontal classes: recall drops to 60.4% while precision remains at 68.0%, yielding an F1-score of 63.9%. This pattern indicates that the model under-predicts leftward gaze rather than over-predicting it: several true left-gaze events are misclassified as other directions, but when the model does predict left, the prediction is correct more often than not. Downward gaze shows the opposite pattern, with the highest precision (86.8%) and a correspondingly high F1-score (80.7%), consistent with vertical gaze cues being more stable across sessions. After calibration with $K = 5$, the precision–recall gap narrows for all directions, and leftward gaze F1 improves from 63.9% to 77.5%.

At the aggregate level, macro F1-score closely tracks balanced accuracy across all evaluation settings (Table 4.3), which reflects the approximately balanced class frequencies in the dataset. The per-class decomposition is therefore more informative than the macro average for diagnosing direction-specific weaknesses.

Table 4.4: Per-class precision, recall, and F1-score under the three main evaluation settings. Values are computed from the aggregated confusion matrices.

Evaluation setting	Direction	Precision	Recall	F1-score
Event-level split	Left	86.1%	85.7%	85.9%
	Right	82.1%	85.5%	83.8%
	Up	83.3%	85.6%	84.4%
	Down	91.3%	85.4%	88.2%
Leave-one-session-out	Left	68.0%	60.4%	63.9%
	Right	66.1%	74.2%	69.9%
	Up	67.5%	75.6%	71.3%
	Down	86.8%	75.5%	80.7%
Calibration-assisted, $K = 5$	Left	77.6%	77.5%	77.5%
	Right	78.6%	76.3%	77.4%
	Up	80.9%	90.8%	85.6%
	Down	91.8%	83.3%	87.4%

4.6.3 Calibration and Angular Error Structure

Table 4.3 shows that a short target-session calibration sequence progressively recovers the cross-session gap, with balanced accuracy rising from 71.5% at $K = 0$ to 82.1% at $K = 5$. The sensitivity to calibration budget, the per-session distribution of gains, and the deployment implications of this trend are analysed in Section 4.8.

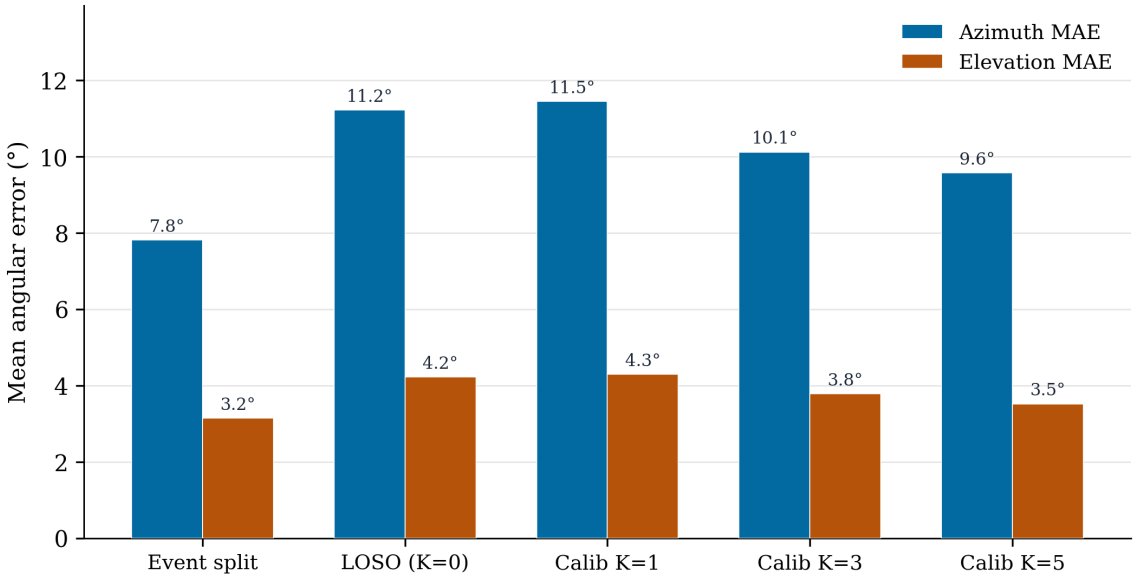


Figure 4.12: Mean azimuth and elevation angular errors across the event-level, leave-one-session-out (LOSO), and calibration-assisted evaluation settings.

A pattern that is visible across all evaluation settings, and that is not fully explained by the calibration analysis, is the persistent asymmetry between azimuth and elevation errors. Figure 4.12 shows that azimuth error is consistently two to three times larger than elevation error, from the event-level split (7.82° versus 3.15°) through to the best calibrated setting (9.58° versus 3.52°). This asymmetry is partly expected because the horizontal gaze range in the stimulus protocol is wider than the vertical range. However, the cross-session increase in azimuth error (+3.41°) is proportionally larger than the elevation increase (+1.07°), and calibration does not close the gap. Horizontal gaze estimation therefore appears more dependent on session-specific radar geometry than vertical estimation, which is consistent with the weaker left-gaze recall observed in the confusion matrix.

4.7 Baseline Comparison

The baseline models defined in Section 3.7.8 provide reference levels for interpreting the leave-one-session-out performance. Table 4.5 reports their results under the identical protocol.

The dominant gap in Table 4.5 lies between the non-temporal and temporal tiers. The best non-temporal classifier reaches 36.1%, only marginally above the 25% chance level, while the SingleStreamCNN reaches 62.5% on the same six channels.

Table 4.5: Baseline comparison under the leave-one-session-out protocol. The temporal models operate on the full $(50, 6)$ event window.

Tier	Model	Params	Bal. acc.	Std (pp)	Az MAE	El MAE
Non-temporal	SVM-RBF	—	36.1%	±8.5	23.1°	6.7°
	Random Forest	—	30.1%	±8.5	26.1°	6.5°
	MLP (2-layer)	—	33.3%	±7.6	23.7°	6.8°
Temporal	SingleStreamCNN	28 k	62.5%	±16.3	16.1°	5.0°
Temporal + dual-stream	DualStreamGazeNet	25 k	71.5%	±14.0	11.2°	4.2°

A 26-point gap cannot be attributed to classifier capacity, since the MLP can already fit complex boundaries in the 36-dimensional feature space. The implication is that temporal ordering within the 50-frame window carries most of the gaze-discriminative information: pooled statistics capture the magnitude of the radar response but discard its direction-specific dynamics, so a leftward and a rightward saccade produce similar amplitude distributions while their temporal sequences differ. The angular error follows the same staircase, with azimuth MAE falling from above 23° at the non-temporal tier to 16.1° once temporal ordering is preserved, and further to 11.2° under the dual-stream design.

The further 9.0-point gain from 62.5% to 71.5% reflects a different mechanism. The two networks receive the same temporal input but differ in three coupled design choices: separation of amplitude and phase-difference channels into independent streams, cross-modal attention before fusion, and a modality-aware hierarchical classification head. The improvement therefore comes from structuring the same information according to its physical origin, not from additional capacity. DualStreamGazeNet contains fewer parameters than SingleStreamCNN (25 k versus 28 k), so the gain is attributable to how the channels are routed, not to a larger model. This matters because the two modalities differ in session sensitivity. Amplitude depends mainly on motion magnitude and is relatively stable across sessions, while phase differences encode geometric path-length relations and shift with face-to-radar geometry. A jointly processed model lacks the structural means to disentangle session-dependent phase shifts from direction-dependent phase patterns. This is also visible in the fold variability, as the SingleStreamCNN shows the highest standard deviation (±16.3 pp), exceeding even the proposed model (±14.0 pp). The individual contribution of each of the three design choices is examined in the ablation study of Section 4.9.

4.8 Calibration Analysis

4.8.1 Cross-Session Variability

The leave-one-session-out mean in Section 4.6 conceals substantial variation between held-out sessions. Figure 4.13 shows balanced accuracy ranging from 41.7% to 93.7%, with some sessions close to within-session performance and others near the lower limit of useful four-class discrimination. This range is too large to be explained by ordinary fold noise. It indicates that the radar gaze signature is highly

session-dependent.
The important interpretation is that low-performing sessions should not be treated as cases where gaze information is absent from the radar signal. The same model performs well on other held-out sessions, and calibration improves most weak sessions. The more plausible direct conclusion is that the feature distribution of some sessions lies outside the support covered by the remaining training sessions. In radar micro-motion sensing, this can occur through modest changes in face-to-radar distance, sensor alignment, head pose, and phase polarity. These mechanisms are discussed further in Chapter 5; the result in this chapter establishes the empirical need for session adaptation or more geometry-robust features.

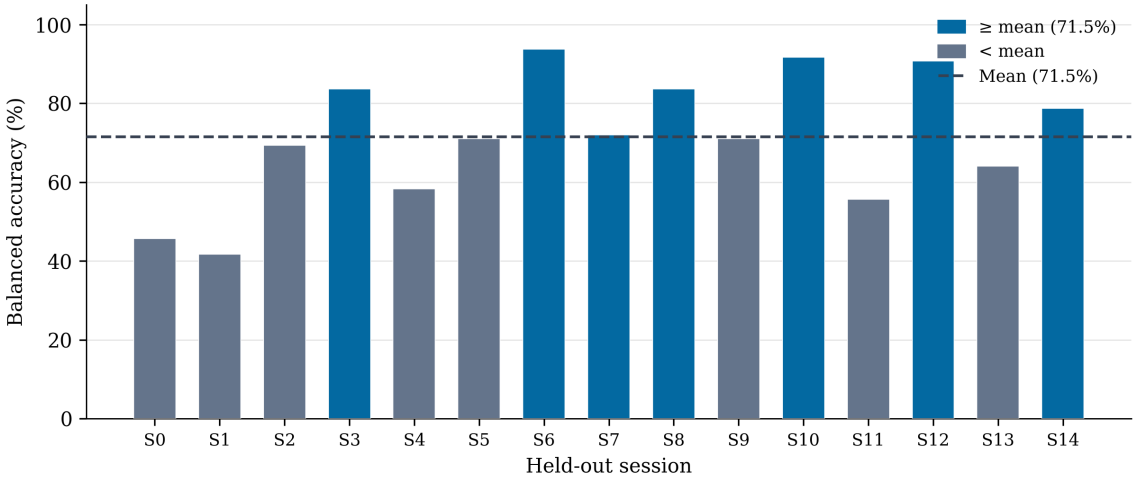


Figure 4.13: Per-session balanced accuracy under the leave-one-session-out protocol. Each bar represents one held-out recording session, and the dashed line marks the cross-session mean.

4.8.2 Sensitivity to Calibration Budget

Few-shot calibration was evaluated by adding K labelled events per gaze direction from the held-out session before testing on the remaining events. The calibration curve in Figure 4.14 shows that the benefit is strongly budget-dependent. The smallest budget, $K = 1$, barely improves classification and does not reduce angular error. This means that calibration is not merely a matter of seeing one target-session example. A single event per direction cannot reliably estimate the session-specific variation in amplitude scaling, phase polarity, event timing, and noise.
The largest efficiency gain occurs between $K = 1$ and $K = 3$. At $K = 3$, balanced accuracy reaches 78.2%, compared with 71.5% without calibration and 82.1% at $K = 5$. Thus 12 labelled events recover a substantial part of the attainable improvement, while the additional eight events from $K = 3$ to $K = 5$ provide a smaller marginal gain. This behaviour suggests diminishing returns; a few examples per class are enough to place the target session closer to the training distribution, but additional examples mainly refine an already useful adaptation.
From a deployment perspective, $K = 3$ and $K = 5$ represent different operating points. $K = 3$ is more efficient, requiring only 12 labelled events, and may be

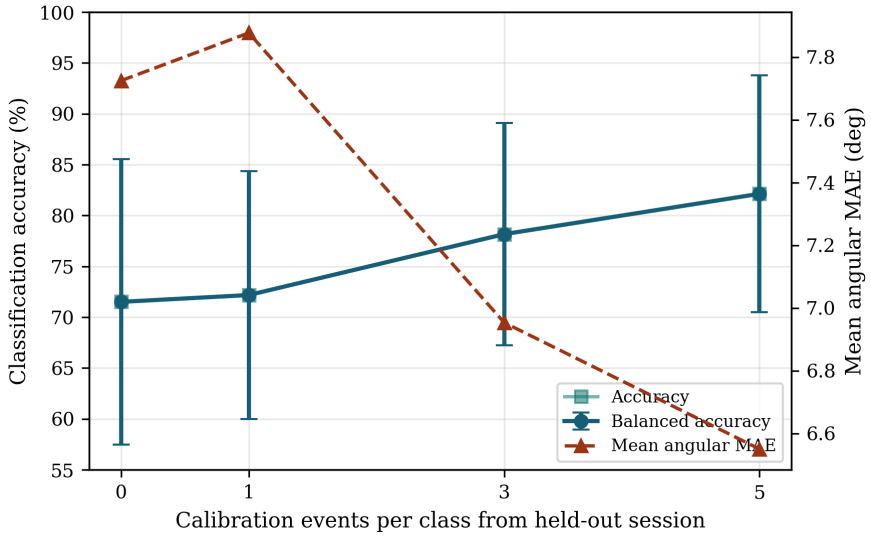


Figure 4.14: Effect of few-shot session calibration on balanced accuracy and mean angular error as a function of the number of calibration events per gaze direction K.

adequate when a short setup procedure is preferred. $K = 5$ requires 20 labelled events and gives the best overall result. The practical constraint is not only the time needed to collect these examples, but also the need to know the intended gaze direction during calibration. This is acceptable in a guided setup with a short visual stimulus sequence, but it remains a limitation for fully unconstrained radar-only use.

4.8.3 Who Benefits from Calibration

Figure 4.15 shows that calibration is most valuable for sessions that perform poorly without it. Sessions with high uncalibrated accuracy improve only slightly, and one already strong session degrades marginally. In contrast, weak sessions tend to move upward after calibration. This asymmetric pattern matters because it distinguishes calibration from a generic regularizer. Calibration does not simply add a uniform accuracy offset. It mainly corrects target sessions whose radar feature distributions are poorly aligned with the source sessions.

This interpretation also explains why calibration does not remove all session variability. If a session is already close to the training distribution, there is little mismatch to correct. If a session is far from the training distribution, a few examples can improve the alignment but may not fully cover all within-session variability. The remaining spread after calibration therefore indicates that few-shot adaptation improves robustness but does not eliminate the need for more invariant sensing and representation design.

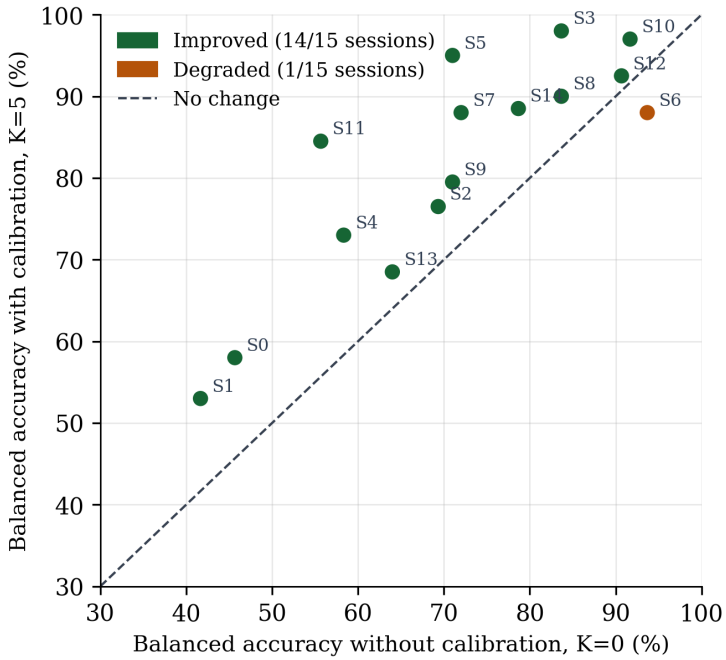


Figure 4.15: Per-session balanced accuracy without calibration ($K = 0$) and with five calibration events per class ($K = 5$). Points above the diagonal indicate improvement after calibration.

4.8.4 Direction-Specific Effects

The calibration effect is direction-dependent. Figure 4.16 shows that leftward gaze is the most affected class under strict leave-one-session-out evaluation and remains the most challenging direction after calibration. At $K = 5$, left-gaze recall improves substantially, from 60.4% to 77.5%, but it does not reach the strongest vertical-class performance. Upward gaze, by contrast, recovers to 90.8%. This difference indicates that calibration does not merely increase all class recalls uniformly.

The direction-specific pattern is important for interpreting the method's effectiveness. Calibration can correct part of the session shift, but the remaining left-gaze weakness suggests that some horizontal-motion cues are more sensitive to sensor geometry than the vertical cues. The result also supports the separate reporting of per-class recall and axis-wise angular error. A single balanced-accuracy value would obscure the fact that the system is more robust for some gaze directions than for others.

4.8.5 Deployment Implications

The calibration study shows a clear tradeoff between session independence and practical reliability. Without calibration, the leave-one-session-out protocol provides the strictest estimate of generalization and exposes the sensitivity of the radar representation to recording geometry. With calibration, the system becomes more reliable in a guided deployment scenario, because a short labelled sequence can substantially reduce the session gap. The key practical question is therefore not whether

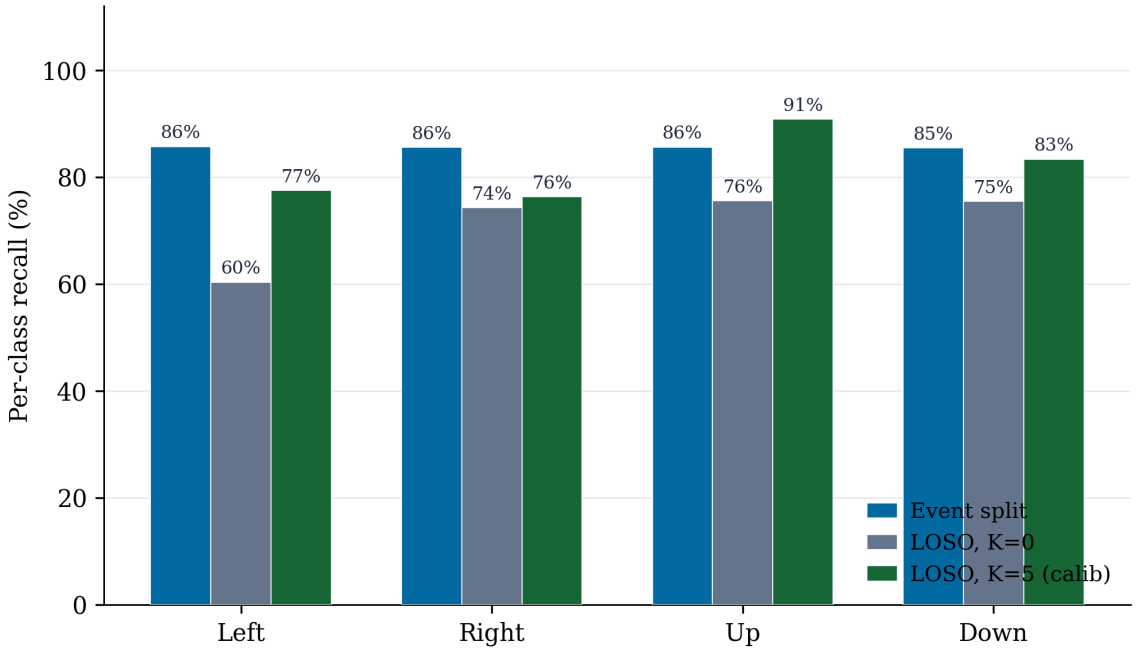


Figure 4.16: Per-direction recall under event-level evaluation, leave-one-session-out (LOSO) evaluation without calibration, and leave-one-session-out (LOSO) evaluation with five calibration events per class.

calibration helps, but how much labelled setup data is acceptable for the target application.

4.9 Ablation Studies

The ablation study identifies which design choices are responsible for the leave-one-session-out performance. Each ablation changes one component while keeping the dataset, evaluation protocol, and remaining training configuration fixed. Figure 4.17 should therefore be read as a contribution analysis: it indicates which components are essential for cross-session robustness and which seemingly reasonable alternatives are harmful.

4.9.1 Radar Feature Contribution

The feature-set ablations confirm that amplitude and phase-difference channels contribute complementary information. Removing the phase-difference channels and retaining only amplitude-related inputs reduces balanced accuracy from 71.5% to 64.6%, a drop of 6.9 percentage points. The converse ablation, removing amplitude channels and retaining only phase differences, produces a larger decrease to 56.5%, a drop of 15.0 percentage points. Neither modality alone reaches the combined performance, and the asymmetry in degradation indicates that amplitude cues carry a larger share of the discriminative information under leave-one-session-out evaluation. This asymmetry is consistent with the physical roles of the two cue families. Amplitude cues describe reflected energy and receiver-dependent strength, providing a

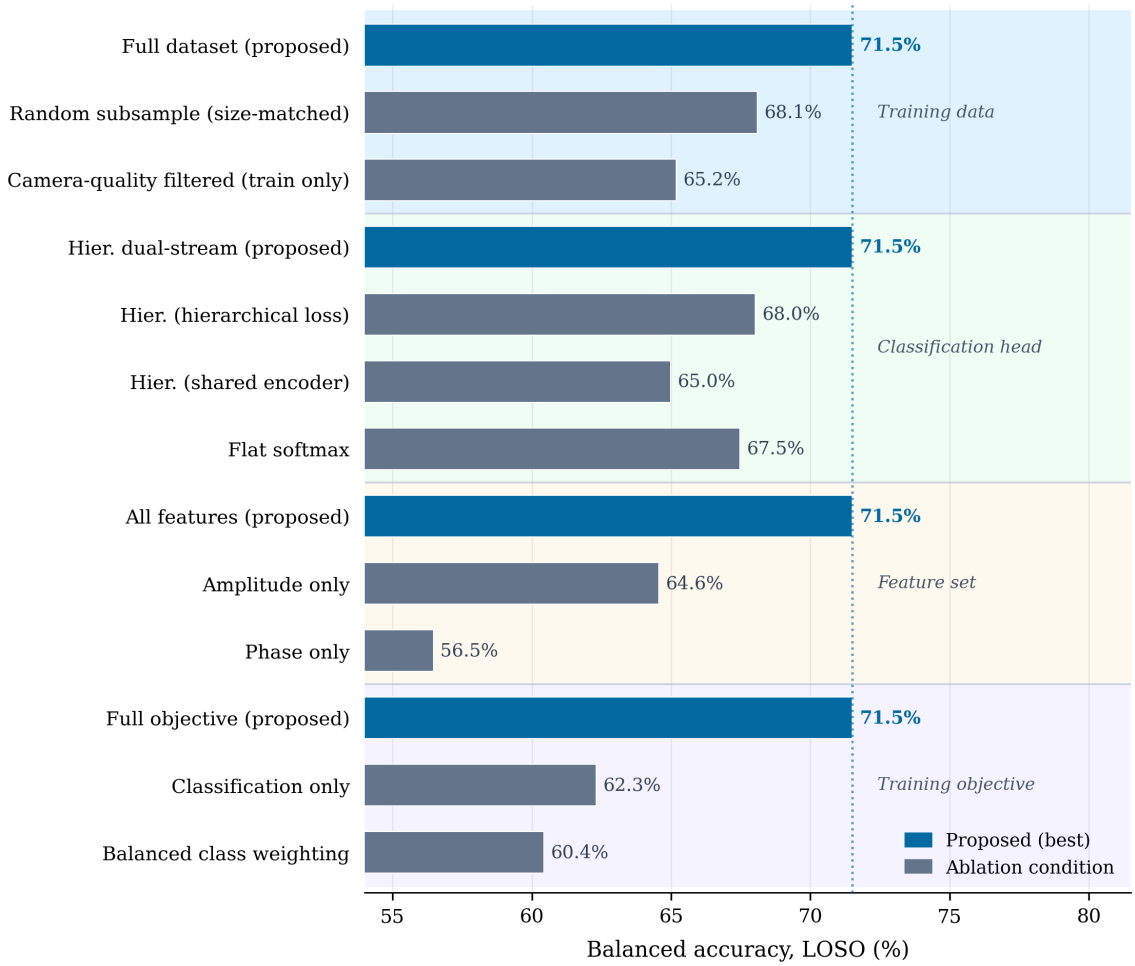


Figure 4.17: Ablation study under the leave-one-session-out (LOSO) protocol. Mean balanced accuracy is shown for the proposed configuration and for ablated alternatives grouped by design axis.

relatively direct measure of motion magnitude. Phase differences provide relative path-length information across receivers but are more sensitive to absolute sensor geometry. Under cross-session evaluation, where face-to-radar distance and alignment vary, the geometric sensitivity of phase features makes them harder to generalize from training sessions alone. Nevertheless, the combined representation substantially outperforms either modality in isolation, confirming that the two cue families are not redundant. The ablation therefore supports both the feature design and the decision to process amplitude and phase cues separately before fusion.

4.9.2 Classification Head Contribution

The head ablations show that the useful contribution is not simply architectural complexity. A flat four-way classifier reaches 67.5%, and non-modality-aware hierarchical variants remain in the 65.0% to 68.0% range, below the proposed 71.5% configuration. A hierarchy by itself is therefore not sufficient. The benefit comes from matching the decision structure to the signal structure.
This is an important result because it connects the model architecture to the sensing problem. The modality-aware hierarchical head allows phase-sensitive information to contribute more directly to horizontal discrimination, while amplitude-related cues remain available for axis and vertical decisions. The gain over both flat and non-modality-aware alternatives suggests that the radar cues are not equally informative for all sub-decisions. The output structure improves performance when it respects this asymmetry.

4.9.3 Data Selection and Objective Contribution

The data-selection ablation shows that a visually cleaner camera trajectory does not necessarily produce a better radar training set. Filtering the training set to retain only events with a clear camera-observed transition reduces balanced accuracy from 71.5% to 65.2%. This result is directly connected to the camera-quality analysis in Section 4.3. The classification target is the controlled stimulus direction, not a thresholded camera label. Removing events because they are ambiguous in the camera trajectory can therefore discard useful radar examples without correcting the class target.
A control experiment isolates the contribution of reduced dataset size from the effect of non-random sample removal. Randomly subsampling the unfiltered training set to the same number of events yields a balanced accuracy of 68.1%, which falls between the full dataset (71.5%) and the camera-quality filtered set (65.2%). Of the total 6.3 percentage-point degradation caused by filtering, approximately half (3.4 pp) is attributable to the reduction in training set size, and the remaining half (2.9 pp) to the non-random nature of camera-quality filtering, which tends to remove events preferentially from sessions with poorer camera conditions and thereby reduces the effective session-level diversity of the training set. The result cautions against using camera quality as a simple proxy for radar training quality. A sample can be ambiguous in the camera-derived trajectory while still carrying useful radar information associated with the controlled stimulus.

The objective ablations identify the auxiliary angle-regression head as one of the strongest contributors. Removing it reduces balanced accuracy to 62.3% and increases azimuth error to 15.85°. This drop shows that the regression head is not merely a secondary output for reporting angular error. The continuous angle targets provide a denser learning signal than four-class labels alone, encouraging the shared encoder to preserve gaze-displacement information that also benefits categorical direction recognition. This is consistent with the general finding that auxiliary tasks can improve shared representations in multi-task learning when the tasks are related [18]. In contrast, class-frequency weighting reduces balanced accuracy to 60.4%. The stimulus protocol produces approximately balanced class frequencies across the full dataset, but the class distribution within individual leave-one-session-out folds can deviate from this overall balance. Inverse-frequency weighting may therefore amplify fold-specific distributional irregularities rather than correct a systematic class imbalance, which could explain both the direction and the magnitude of the observed degradation.

4.9.4 Design Implications

Across the ablations, the most important design principle is that performance improves when the model respects the structure of the sensing problem. Amplitude and phase-difference features each carry information that the other lacks, with amplitude contributing more to overall cross-session robustness and phase differences providing complementary geometric cues that improve the combined representation. The modality-aware hierarchical head uses this information in the sub-decisions where it is most useful. The full event set benefits both from its larger size and from its broader session-level coverage, as confirmed by the size-matched random subsampling control. The auxiliary regression objective provides a denser learning signal that encourages the encoder to preserve continuous gaze displacement. Conversely, stricter camera-based filtering and class-frequency weighting reduce performance, suggesting that neither camera-trajectory ambiguity nor class imbalance is a dominant limitation in this dataset.

The ablation results therefore refine the interpretation of the main performance results. The proposed system does not work simply because it has a neural network with multiple outputs. Its strongest components are those that align the learning problem with the physics of the radar signal and with the structure of the supervision. This conclusion provides the basis for the broader discussion of robustness, calibration, and deployment limitations in Chapter 5.

5

Discussion

5.1 Interpretation of Results

The results in Chapter 4 support the central premise of this thesis: short event-centred mmWave radar windows contain measurable information about gaze direction. Under the event-level split, the model achieves high four-direction classification accuracy using only radar-derived amplitude and phase features as input. The camera serves as a label source during training and evaluation but is not part of the inference path. This separation is important because it shows that the radar signal around a gaze event carries learnable structure that a temporal model can map to the intended gaze direction, rather than merely correlating with unrelated facial motion or session-specific artefacts.

However, the leave-one-session-out results reveal a clear limitation: performance degrades substantially on unseen sessions, and the gap relative to the event-level split is too large to attribute to validation noise. The degradation concentrates on horizontal gaze, with leftward gaze recall dropping more than other directions and azimuth error increasing more than elevation error. Since all four directions achieve high accuracy under the event-level split, the weakness is not intrinsic to leftward gaze but reflects how horizontal gaze interacts with the recording geometry.

Horizontal gaze shifts likely depend more on lateral displacement and the purely horizontal receive baseline (RX2–RX0), which is sensitive to sensor placement, whereas vertical shifts produce a stronger signature in the diagonal baseline (RX1–RX0) and in amplitude-based eyelid cues that are less dependent on receiver geometry. This would explain why vertical gaze recall is more robust across sessions and why the event-level feature traces in Figure 4.8 show clearer vertical directional contrast in the diagonal phase-difference channel (b) than horizontal contrast in the horizontal-baseline channel (c). The interpretation is also consistent with the importance of phase-difference features in the ablation study. The underlying issue is that small changes in sensor-to-face geometry between sessions alter the amplitude and phase patterns associated with the same eye movement, and the learned representation is not yet invariant to these changes.

The ablation study identifies which components contribute most to cross-session robustness. The largest single-component effect comes from the auxiliary regression objective: removing it and training with classification alone reduces balanced accuracy from 71.5% to 62.3%. The regression loss constrains the shared representation to encode the continuous magnitude and axis of gaze displacement, and this additional structure evidently helps the classifier distinguish directions more reliably

across sessions. Removing the phase-difference channels causes the second-largest drop (71.5% to 64.6%), confirming that amplitude alone does not capture all discriminative information and that inter-receiver phase provides geometrically distinct cues that complement amplitude-based features. On the classification-head side, the modality-aware hierarchical design outperforms both a flat classifier (67.5%) and non-modality-aware hierarchical variants (65.0–68.0%), indicating that performance depends not just on architectural complexity but on matching the decision structure to the signal: horizontal and vertical gaze decisions draw on amplitude and phase in different proportions.

The camera-derived ground truth is adequate for supervised learning under the controlled protocol, with angular errors on the order of a few degrees, but it is not a calibrated clinical eye tracker. Its residual cross-axis components and location-dependent bias justify the two-tier label design used in this thesis and mean that small differences in regression error should not be over-interpreted.

Calibration bridges the gap between controlled feasibility and practical deployment. With three labelled events per direction, balanced accuracy rises from 71.5% to 78.2%; with five it reaches 82.1%. That a small amount of target-session data recovers most of the cross-session loss indicates that the dominant error source is distributional shift in amplitude and phase statistics rather than a fundamental failure of the learned features. Three to five events per direction thus provide a practical operating point at modest annotation cost.

5.2 Practical Implications for Driver Monitoring

The most realistic near-term role for radar gaze estimation in driver monitoring is as a complement to camera-based tracking rather than a replacement. Camera systems remain stronger for dense spatial gaze estimation when the face is visible, but they are affected by illumination changes and raise privacy concerns because they capture identifiable facial imagery. Radar operates independently of ambient light and does not produce visual images, making it useful when lighting is poor, when image privacy is a concern, or when camera confidence is reduced. A multimodal system could use camera-based gaze under reliable visual conditions and radar-derived gaze events as a secondary signal for robustness or privacy-sensitive operation.

The event-centred nature of the present system is important for practical interpretation. The model estimates gaze direction and angle for detected eye-movement intervals rather than producing dense frame-by-frame gaze at every moment. This is compatible with some driver monitoring tasks, such as detecting glance shifts toward mirrors, instrument clusters, or off-road regions, where gaze transitions are meaningful events. It is less directly suited to applications requiring continuous point-of-regard estimation on a display or detailed fixation tracking. For driver monitoring, this distinction suggests that radar gaze estimation may be most useful as an event detector and directional estimator, especially when combined with temporal logic that accumulates evidence over multiple glances.

Sensor placement is likely to be a decisive practical factor. The results show that session geometry strongly affects performance, and a vehicle cabin would introduce additional variation through steering wheel position, seat adjustment, driver height,

head pose, and vibration. A radar mounted near the dashboard or steering column would observe the face from a fixed but driver-dependent angle. The receiver geometry that supports phase-difference features in the laboratory may not transfer directly to all in-vehicle placements. For practical use, the radar position, field of view, range-bin selection, and calibration procedure should be designed together rather than treated as independent engineering choices.
Real-time use also requires a reliable radar-only event proposal stage. The model evaluated in the main experiments uses event-centred windows, and Chapter 3 describes a radar-only inference pipeline in which candidate eye-movement intervals are detected before gaze estimation. In a deployed driver monitoring system, this event proposal stage must operate continuously and reject non-gaze motion such as head turns, facial expressions, vibration, or hand movement near the face. The present work demonstrates the feasibility of classifying controlled event windows, but a full driver monitoring system would require end-to-end validation of event detection, window construction, prediction latency, and false-positive behaviour in continuous recordings.

5.3 Limitations

The first limitation is dataset scale and diversity. The experiments are based on a limited number of recording sessions collected under controlled conditions. Although leave-one-session-out evaluation provides a useful estimate of cross-session generalization, it does not fully represent the variability expected across a large population of drivers, facial anatomies, eyewear conditions, head poses, seating positions, and sensor placements. The high variance across held-out sessions already indicates that the current dataset does not cover all relevant sources of variation. Stronger conclusions about generalization require a larger and more diverse dataset with explicit subject-level and environment-level variation.
The second limitation is the controlled laboratory protocol. The stimulus-driven four-direction task provides clean labels and repeatable gaze events, which is valuable for method development. However, real driver gaze behaviour is less constrained. Glances vary in amplitude, duration, starting point, and target location. Drivers may move their heads, blink, speak, or interact with the vehicle while shifting gaze. The present protocol isolates gaze-related radar cues, but it does not fully capture the complexity of in-vehicle behaviour. The reported results should therefore be interpreted as evidence of controlled feasibility rather than as a direct estimate of road-ready performance.
The camera-derived ground truth is another limitation. As discussed in Section 5.1, the MediaPipe-based reference is adequate for supervised learning but remains a monocular landmark-based approximation with location-dependent error. This uncertainty affects the regression targets directly, and classification results based on stimulus labels are therefore more robust to camera reference noise than regression results based on camera-derived angles.
Calibration-assisted results also require careful interpretation. They show that a small amount of target-session data can improve performance, but they do not represent strict radar-only deployment in the strongest sense. The calibration ex-

amples require known direction labels, which in practice would need to come from a guided setup sequence, a display protocol, or another trusted reference. Calibration also changes the evaluation condition by allowing target-session information into training. This is useful for estimating session-adaptation benefit, but it should be reported separately from leave-one-session-out performance, as done in Chapter 4.
The present model is also limited by its event-level formulation. The network assumes a fixed-length window around a candidate gaze event, and the radar-only inference pipeline requires a separate event proposal stage that has not been validated under continuous real-world motion. Errors in event detection, anchor timing, or range-bin selection would propagate into the gaze estimator, so the reported performance should not be interpreted as a continuous tracking estimate. Relatedly, the analysis of radar geometry remains indirect: the current experiments do not systematically vary radar position, head pose, or eye-to-sensor distance, so the physical explanation for direction-specific errors cannot be fully separated from dataset-specific effects.

5.4 Future Work

Expanding the dataset could be the first priority. A larger multi-subject dataset should include more recording sessions, a wider range of facial geometry, eyewear conditions, head poses, and natural gaze amplitudes. It should also separate within-subject, cross-session, and cross-subject evaluation protocols. This distinction is important because a model that adapts well across sessions for the same subject may still fail to generalize across subjects. A larger dataset would also make it possible to evaluate whether the left-gaze weakness observed here is a stable geometric effect or a property of the current recording setup.
The next step is in-vehicle validation. Laboratory recordings are useful for isolating the signal, but driver monitoring requires robustness to seat position, vehicle vibration, changing illumination, head motion, and natural driving behaviour. An in-vehicle dataset should include both guided calibration sequences and more natural driving or simulator tasks. Such data would allow the radar-only event detector, feature-window adapter, and gaze estimator to be evaluated as a complete pipeline rather than as separate components.
Improved calibration and adaptation methods are also needed. The present results show that three to five labelled events per direction can substantially reduce session shift. Future work should examine whether this process can be shortened, made unsupervised, or replaced by self-supervised session adaptation. Possible directions include estimating session-specific radar normalization from a neutral baseline, adapting phase polarity and amplitude scale without labelled gaze events, or using confidence-based online adaptation during repeated glances. The practical goal is to reduce the dependence on a guided labelled calibration sequence while retaining the robustness benefits observed in this thesis.
Future models should also make better use of radar structure. The present system uses a selected range bin and a compact set of amplitude and phase-difference features. More robust models could incorporate multiple range bins, range-angle representations, or explicit uncertainty estimates for event timing and range selection.

Architectures that combine temporal attention with physics-informed constraints may help separate eye motion from head motion and facial motion. Additional work should also examine whether the modality-aware hierarchy can be extended to more gaze regions or continuous angular sectors without losing the interpretability of the current four-direction structure.
Finally, sensor fusion and real-time validation are needed to move toward deployment. Rather than treating radar and camera as competing modalities, future systems may combine them according to confidence and context, fusing at the level of event detection, feature representation, or decision confidence. A deployable system must also process continuous radar streams end to end, rejecting non-gaze motion within acceptable latency, and report metrics beyond classification accuracy, including event-detection recall, false-positive rate, and robustness to motion artefacts.

6
Conclusion

This thesis investigated whether 60 GHz FMCW radar can serve as a sensing modality for four-direction gaze event classification and continuous angle regression from short event-centred windows. The experimental results confirm that the radar signal around a voluntary eye-movement event contains learnable structure specific to the intended gaze direction, and that a compact temporal model operating on radar-derived features alone can exploit this structure.

The approach converts raw FMCW frames to the range domain through per-chirp spectral processing, averages the complex response at a session-adaptive range bin across chirps, and constructs a six-channel feature representation from the resulting frame-level samples: three window-normalised receiver amplitudes, two inter-receiver phase differences, and one inter-receiver amplitude difference. This design reflects a physical asymmetry: amplitude channels capture eyelid-related reflectivity changes most relevant to vertical gaze, while phase differences cancel common-mode head-motion drift and encode lateral wavefront shifts associated with horizontal gaze. The ablation analysis confirms that removing phase-difference channels degrades cross-session accuracy more than any other single modification.

DualStreamGazeNet extends this amplitude-phase separation into the network architecture through distinct convolutional encoders, cross-modal attention fusion, a modality-aware hierarchical classifier, and an auxiliary regression head. The ablation study shows that each element contributes to session-robust performance by margins larger than would be expected from comparable changes in parameter count, indicating that architectural choices informed by signal physics yield greater robustness than increases in model capacity alone.

Session variability is the dominant constraint. Under leave-one-session-out evaluation, performance degrades substantially, with horizontal gaze affected more than vertical gaze. A calibration sequence of three to five labelled events per direction recovers much of the lost accuracy, confirming that the cross-session gap is primarily a geometry-driven distribution shift rather than a fundamental signal limitation. The radar-only inference pipeline further demonstrates that candidate eye-movement intervals can be proposed from continuous radar data through motion-energy thresholding, without camera or stimulus input at inference time.

The main boundary condition is the controlled laboratory setting with fifteen sessions from two participants and a four-direction stimulus protocol. The path to in-vehicle deployment requires larger and more diverse datasets, improved session-adaptation methods, validated event detection in continuous operation, and integration with complementary sensing modalities. The present results provide a technical foundation and a set of physics-informed design principles for that effort.

References

[1] P. K. Sharma and P. Chakraborty, "A review of driver gaze estimation and application in gaze behavior understanding," *Eng. Appl. Artif. Intell.*, vol. 133, no. PB, Jul. 2024, ISSN: 0952-1976. DOI: 10.1016/j.engappai.2024.108117. [Online]. Available: https://doi.org/10.1016/j.engappai.2024.108117.

[2] Y. Cheng et al., "Appearance-based gaze estimation with deep learning: A review and benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7509–7528, 2024. DOI: 10.1109/TPAMI.2024.3393571. [Online]. Available: https://doi.org/10.1109/TPAMI.2024.3393571.

[3] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010. DOI: 10.1109/TPAMI.2009.30. [Online]. Available: https://doi.org/10.1109/TPAMI.2009.30.

[4] A. Soumya, C. K. Mohan, and L. R. Cenkeramaddi, "Recent advances in mmwave-radar-based sensing, its applications, and machine learning techniques: A review," *Sensors*, vol. 23, no. 21, p. 8901, 2023. DOI: 10.3390/s23218901. [Online]. Available: https://doi.org/10.3390/s23218901.

[5] J. Hu et al., "BlinkRadar: Non-intrusive driver eye-blink detection with UWB radar," in *Proceedings of the 42nd IEEE International Conference on Distributed Computing Systems*, 2022, pp. 1040–1050. DOI: 10.1109/ICDCS54860.2022.00104. [Online]. Available: https://doi.org/10.1109/ICDCS54860.2022.00104.

[6] D. Patscheider, R. Wu, A. Broquetas, A. Aguasca, and J. Romeu, "Eyelid dynamics characterization with 120 ghz mmw radar," *Sensors*, vol. 24, no. 23, p. 7464, 2024. DOI: 10.3390/s24237464. [Online]. Available: https://doi.org/10.3390/s24237464.

[7] S. Zhang, Q. Wang, K. Song, Q. Yan, and H. Zeng, "RadEye: Tracking eye motion using FMCW radar," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, New York, NY, USA: Association for Computing Machinery, 2025, pp. 1–13. DOI: 10.1145/3706598.3713775. [Online]. Available: https://doi.org/10.1145/3706598.3713775.

[8] M. A. Richards, *Fundamentals of Radar Signal Processing*, 2nd ed. New York: McGraw-Hill Education, 2014, ISBN: 9780071798327. [Online]. Available: https://www.fccdecastro.com.br/pdf/FRSPMR.pdf.

[9] V. C. Chen, *The Micro-Doppler Effect in Radar*, 2nd ed. Norwood, MA: Artech House, 2019, ISBN: 9781630815486. [Online]. Available: http://ieeexplore.ieee.org/document/9098778.

[10] I. Bekerman, P. Gottlieb, and M. Vaiman, "Variations in eyeball diameters of the healthy adults," *Journal of Ophthalmology*, vol. 2014, p. 503 645, 2014. DOI: 10.1155/2014/503645. [Online]. Available: https://doi.org/10.1155/2014/503645.

[11] A. T. Bahill, M. R. Clark, and L. Stark, "The main sequence, a tool for studying human eye movements," *Mathematical Biosciences*, vol. 24, no. 3–4, pp. 191–204, 1975. DOI: 10.1016/0025-5564(75)90075-9. [Online]. Available: https://doi.org/10.1016/0025-5564(75)90075-9.

[12] R. J. Leigh and D. S. Zee, *The Neurology of Eye Movements*, 5th ed. New York, NY, USA: Oxford University Press, 2015, ISBN: 9780199969289. DOI: 10.1093/med/9780199969289.001.0001.

[13] C. Evinger, K. A. Manning, and P. A. Sibony, "Eyelid movements. mechanisms and normal data," *Investigative Ophthalmology & Visual Science*, vol. 32, no. 2, pp. 387–400, 1991. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1993591/.

[14] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, 2019. [Online]. Available: https://research.google/pubs/real-time-facial-surface-geometry-from-monocular-video-on-mobile-gpus/.

[15] Google AI Edge, *MediaPipe Iris*, https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/iris.md, Accessed 2026-05-12, 2024.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018. [Online]. Available: https://arxiv.org/abs/1803.01271.

[17] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need.

[18] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997. DOI: 10.1023/A:1007379606734. [Online]. Available: https://doi.org/10.1023/A:1007379606734.

[19] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964. DOI: 10.1214/aoms/1177703732. [Online]. Available: https://doi.org/10.1214/aoms/1177703732.

[20] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1–2, pp. 31–72, 2011. DOI: 10.1007/s10618-010-0175-9. [Online]. Available: https://doi.org/10.1007/s10618-010-0175-9.

[21] Y. Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016. [Online]. Available: https://www.jmlr.org/papers/v17/15-239.html.

[22] L. Ma, Y. Ye, C. Gu, and J. Mao, "High-accuracy contactless detection of eyes' activities based on short-range radar sensing," in *2022 IEEE MTT-S International Microwave Biomedical Conference*, ser. IMBioC, 2022. DOI: 10.1109/IMBIoC57193.2022.9790181. [Online]. Available: https://ieeexplore.ieee.org/document/9790181/.

[23] C. Xu, X. Zheng, Z. Ren, L. Liu, and H. Ma, "UHead: Driver attention monitoring system using UWB radar," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, 2024. DOI: 10.1145/3643551. [Online]. Available: https://doi.org/10.1145/3643551.

[24] J. Jung, J. Kim, S.-C. Kim, and S. Lim, "Eye-gaze tracking based on head orientation estimation using FMCW radar sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, 2024. DOI: 10.1109/TIM.2024.3472779. [Online]. Available: https://ieeexplore.ieee.org/document/10723245/.

[25] R. Ma, Y. Morimoto, J. S. Ho, S. Shiu, and J. Zhu, "mmET: Mmwave radar-based eye tracking on smart glasses," in *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, 2025. DOI: 10.1145/3715014.3722050. [Online]. Available: https://doi.org/10.1145/3715014.3722050.

[26] S. S. Healy and C. N. Stephan, "Mean human corneal diameter and palpebral fissure lengths as scales for forensic analysis of photographed faces: An analytical review," *International Journal of Legal Medicine*, 2026. DOI: 10.1007/s00414-026-03733-0. [Online]. Available: https://doi.org/10.1007/s00414-026-03733-0.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[29] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx.

[30] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2, Morgan Kaufmann, 1995, pp. 1137–1143. [Online]. Available: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf.

[31] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proceedings of the 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764. [Online]. Available: https://doi.org/10.1109/ICPR.2010.764.

A
Appendix

A.1 Supplementary radar–gaze overlays

This appendix extends the qualitative discussion of Section 4.4 with additional sessions. For each example, the upper subplot shows phase-derived displacement together with camera-based gaze angle, and the lower subplot shows logarithmic amplitude with the same gaze trace on a common time axis. Range-bin indices and viewer processing annotations appear in the embedded figure titles and may differ across recordings; the purpose here is to show that amplitude and phase tracks remain temporally aligned with oculomotor transitions under typical desktop geometry, not to fix a single numerical pipeline for all sessions.

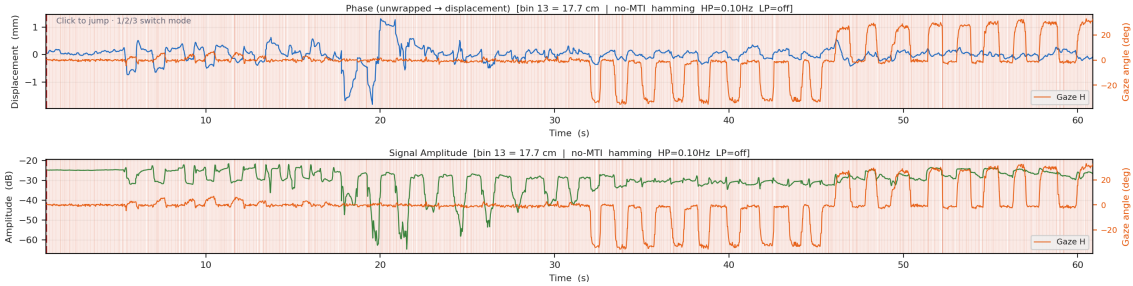


Figure A.1: Session A, horizontal gaze (*Gaze H*).

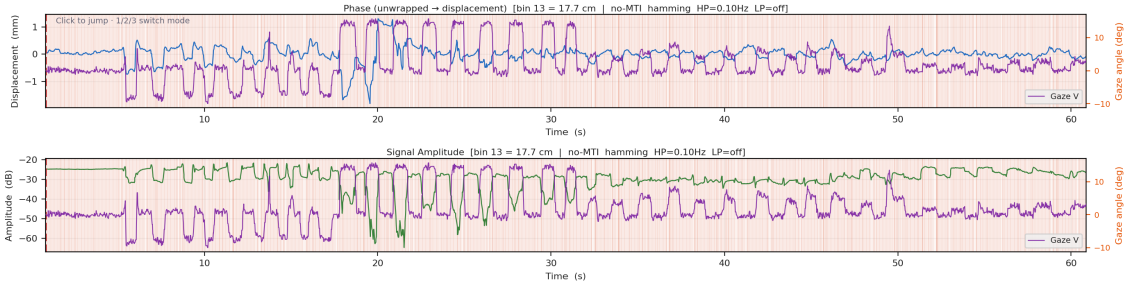


Figure A.2: Session A, vertical gaze (*Gaze V*).

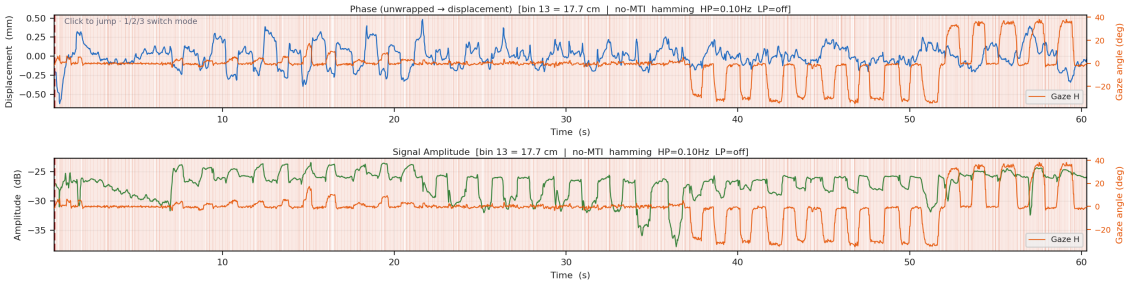


Figure A.3: Session B, horizontal gaze (*Gaze H*).

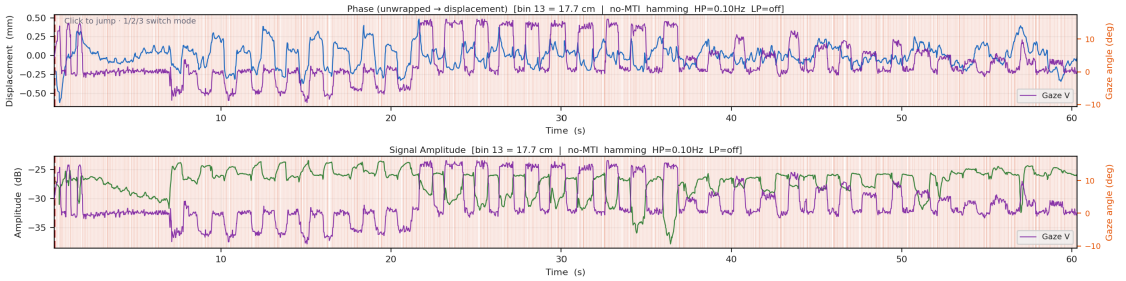


Figure A.4: Session B, vertical gaze (*Gaze V*).

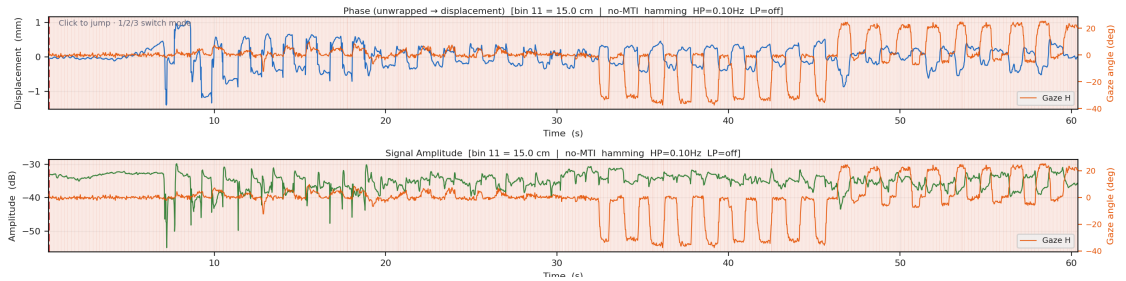


Figure A.5: Session C, horizontal gaze (*Gaze H*).

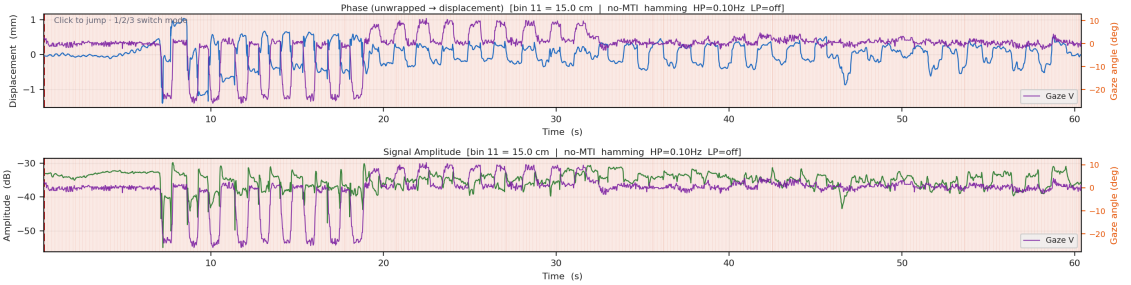


Figure A.6: Session C, vertical gaze (*Gaze V*).

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY