

Pixel-Level Registration Method and Dataset Construction for Multimodal Image Data in Intelligent Cockpits

Xin Ning^{1,2,3}, Kai Zhao^{1,2}, Dan Li^{1,2}, Jin Ning^{1,4,5}, Weijun Li^{1,3,6,7}, and Baoli Lu^{1,*}

Abstract Driver behavior monitoring is fundamental to safety assurance and interaction optimization in intelligent cockpits. To address key challenges in existing research, including limited modality diversity, cross-view spatial misalignment, and insufficient robustness of multimodal perception, we construct a multimodal intelligent-cockpit image dataset consisting of synchronously captured RGB, infrared (IR), and depth images. We also propose a two-stage registration method for dataset preprocessing that combines geometric mapping and mask-guided inpainting to achieve pixel-level alignment across modalities. Using the constructed dataset, we evaluate the impact of different modality combinations on object detection. Results on two baseline multimodal detectors show that the tri-modal setting significantly outperforms RGB-only input in both accuracy and robustness, particularly under challenging lighting conditions, with mAP@0.5:0.95 gains of 36.8% and 36.3%, respectively. These findings verify the effectiveness of the proposed dataset for multimodal object detection and driver behavior understanding. The proposed dataset provides a valuable benchmark for multimodal perception in intelligent cockpits and facilitates future research on driver monitoring and active safety.

Keywords intelligent cockpit; dataset; multimodal perception; pixel-level registration; object detection

1 Introduction

With the rapid advancement of intelligent transportation systems and autonomous driving technologies, accurate percep-

tion and analysis of driver states and behaviors have become a critical component for ensuring road safety and improving the driving experience. According to the *Global Status Report on Road Safety 2023* published by the World Health Organization [1], road traffic accidents remain a major threat to public health worldwide, causing approximately 1.19 million deaths each year and ranking as the leading cause of death among individuals aged 5–29 years. Studies have shown that more than 70% of road traffic accidents are associated with unsafe or risky driving behaviors [2]. For example, smoking while driving not only distracts the driver and increases the risk of one-handed driving, but may also cause safety-critical incidents due to falling ash or dropped cigarette butts, thereby threatening driving safety. Therefore, developing systems capable of monitoring driver behaviors accurately and in real time is of great practical significance for active safety warning and accident prevention.

At present, most vision-based driver behavior detection methods rely primarily on the RGB modality. Although these methods have achieved promising results under favorable illumination conditions, their performance is often severely degraded in complex and dynamic real-world driving scenarios. For instance, under non-ideal lighting conditions such as nighttime, tunnels, backlighting, or strong glare, RGB images may suffer from underexposure or overexposure, loss of detail, and color distortion, resulting in a substantial decline in the detection accuracy of small objects such as hands and cigarettes. To overcome the limitations of single-modality perception, multimodal fusion has emerged as an effective solution. By integrating complementary information from different sensors, the robustness and accuracy of detection models can be significantly improved. For example, IR data remain informative under low-light conditions and can compensate for the information loss of RGB images caused by insufficient illumination, while depth images provide three-dimensional spatial cues that help distinguish targets from the background and mitigate the effects of occlusion. Existing studies in multispectral object detection have demonstrated that fusing RGB and IR information can effectively improve robustness against modality discrepancies and noise, yielding superior performance in pedestrian detection [3] and envi-

¹ Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China.

² College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China.

³ School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing 101408, China.

⁴ Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.

⁵ School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.

⁶ School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China.

⁷ Zhongguancun Academy, Beijing 100094, China.

* Corresponding author: Baoli Lu (lubao@semi.ac.cn)

Manuscript received: 20xx-xx-xx; revised: 20xx-11-xx; accepted: 20xx-00-xx



ronmental monitoring [4]. Likewise, the fusion of RGB and depth information has also been shown to provide richer features and improve detection performance in general object detection tasks [5, 6].

However, the application of multimodal fusion to driver behavior analysis still faces two major challenges. The first challenge lies in cross-modal registration and alignment. Sensors of different modalities usually have different fields of view, resolutions, and imaging mechanisms. Direct fusion of unaligned images leads to feature mismatch and seriously degrades model performance. Existing studies have emphasized that accurate pixel-level alignment is a prerequisite for effective utilization of multimodal information, yet many currently available datasets lack such high-quality aligned data, and this limitation has become a major obstacle to multimodal perception research [7]. The second challenge is the absence of high-quality, large-scale, scenario-specific benchmark datasets. Just as the emergence of large-scale annotated datasets such as SUN RGB-D [8] substantially advanced general scene understanding, driver behavior analysis also urgently requires a dedicated multimodal dataset for complex in-cabin environments. Such a dataset should provide precise pixel-level registration and rich annotations to support the training and evaluation of data-driven deep learning models. At present, most publicly available datasets are limited to a single modality or suffer from insufficient temporal synchronization and spatial alignment accuracy across modalities, making them inadequate for robust detection and fine-grained modeling of subtle driver behaviors such as smoking.

To address these challenges, this paper focuses on the construction and application of a multimodal driver behavior image dataset for intelligent cockpit scenarios. We construct an intelligent-cockpit multimodal image dataset, the Cockpit Multimodal Imaging Dataset (CMMID), and establish a two-stage pixel-level registration method, the Cockpit Multimodal Imaging Registration (CMMIR). The main contributions of this paper are summarized as follows:

(1) We construct a large-scale multimodal intelligent-cockpit image dataset, CMMID, featuring day–night coverage, multiple object categories, and pixel-level aligned multimodal samples. The dataset contains 17,343 synchronized RGB-IR-depth image triplets and more than 80,000 instance annotations for seven categories, including the steering wheel, safety belt, and mobile phone, thereby providing a valuable benchmark for downstream tasks such as multimodal fusion and driver monitoring systems (DMS).

(2) To address the registration requirements of multimodal object detection datasets and the hole artifacts caused by

disparity, we construct a two-stage pixel-level registration method, CMMIR, which combines geometric mapping with mask-guided image inpainting. In the inpainting stage, a cross-modal conditional guidance module and a frequency-domain decoding module are introduced to improve semantic consistency and structural continuity.

(3) Under a unified training and evaluation protocol, we systematically compare the performance differences between multimodal and single-modal models. Using the same training paradigm, we investigate the detection performance of RGB, RGB + IR, RGB + depth, and RGB + IR + depth settings, thereby validating the effectiveness of the constructed dataset for multimodal object detection.

2 Related Work

2.1 Advances in Cross-Modal Image Registration

Cross-modal image registration aims to establish dense spatial correspondences between images acquired by different sensors but depicting the same scene, and serves as a key pre-processing step for multi-source information fusion. The core challenge of this task lies in overcoming significant geometric deformation and nonlinear radiometric intensity differences across modalities. According to the evolution of technical routes, existing cross-modal registration methods can generally be divided into intensity-based methods, feature-based methods, and deep-learning-based methods.

Intensity-based registration methods directly optimize image alignment using pixel values. Among them, mutual information is one of the most representative and widely used strategies. Originating from information theory, this method aligns two images by maximizing the statistical dependency between their intensity distributions, without imposing any prior assumption on the brightness relationship between modalities, such as linearity. This property makes it naturally suitable for multimodal image registration. Early pioneering works, such as those by Maes *et al.* and Pluim *et al.*, systematically introduced mutual information into medical image registration and verified its effectiveness and robustness for aligning CT, MRI, PET, and other multimodal images [9, 10]. However, traditional mutual-information-based methods usually rely on a global metric, are computationally expensive, and exhibit many local optima in the objective function, making them sensitive to initialization. These limitations restrict their application in scenarios requiring fast estimation of dense deformation fields.

Unlike intensity-based approaches, feature-based methods first detect stable and repeatable salient structures in images, such as corners and edges, and then estimate transformations

by matching the corresponding descriptors. The fundamental theory of geometric transformation is comprehensively presented in the classical work of Hartley *et al.* [11]. Traditional handcrafted features such as SIFT are invariant to linear illumination changes, but they struggle to cope with radiometric discrepancies across modalities. To address this issue, various descriptors tailored to cross-modal scenarios have been proposed. For example, RIFT [12] and its improved version RIFT2 [13] employ phase congruency instead of gradient information to construct a maximum index map descriptor, substantially improving matching robustness in remote-sensing scenarios with severe radiometric differences, such as SAR–optical and IR–optical image pairs. With the development of deep learning, learning modality-invariant feature descriptors with neural networks has become a new research trend. Heinrich *et al.* proposed the MIND descriptor [14], which captures self-similarity within local neighborhoods and therefore does not depend on specific pixel intensities, making it naturally suitable for multimodal settings. Building on this idea, Achermann *et al.* further proposed the MultiPoint framework [15], which adopts a multispectral homography-adaptive strategy and trains an end-to-end deep network to jointly detect keypoints and generate descriptors, achieving excellent performance on thermal–optical aerial image registration tasks. These methods successfully combine the strong feature extraction capability of deep learning with the traditional feature matching pipeline, but their performance is still constrained by the accuracy and sparsity of keypoint detection. In recent years, end-to-end deep learning methods have attempted to use a unified network to directly regress transformation parameters or dense displacement fields between images, completely bypassing the separate steps of feature detection and matching. According to the regression target, these methods can be divided into parametric models and non-parametric models. Parametric models mainly estimate global low-degree-of-freedom transformations, such as homography matrices. Le *et al.* proposed a multi-scale VGG-style convolutional network that takes image pairs as input and progressively estimates and refines the homography transformation, thereby enabling end-to-end global geometric regression [16]. This approach converts a complex geometric estimation problem into a supervised regression task and demonstrates the great potential of deep learning for solving traditional geometric vision problems. Non-parametric models, in contrast, aim to estimate pixel-level dense correspondences, namely optical flow or deformation fields. RAFT [17] constructs a 4D correlation volume over all pixel pairs and iteratively optimizes the optical flow field with a recur-

rent neural network. This “extract–query–update” paradigm substantially improves the accuracy and generalization ability of optical flow estimation and sets a new benchmark for dense registration tasks.

Although not directly designed for registration, advanced generative models have shown strong capabilities in related fields such as image inpainting, thereby providing new ideas for addressing the challenges that arise in cross-modal registration. For example, Suvorov *et al.* used fast Fourier convolution to inpaint large missing regions, demonstrating the effectiveness of frequency-domain analysis for capturing global image context [18]. Similarly, models based on Transformers [19] and partial convolutions [20] can generate semantically coherent content by leveraging powerful long-range dependency modeling. More recent generative models, such as denoising diffusion probabilistic models (DDPMs) [21], can produce highly realistic and diverse image content. The core advantage of these techniques lies in their strong capability for image understanding and content generation.

Traditional intensity-based and geometric parametric methods provide a solid theoretical foundation for accurate alignment, while feature-based methods and end-to-end deep learning models have made significant progress in handling severe radiometric differences and estimating dense correspondences. At the same time, advanced generative inpainting techniques, especially Transformer-based and diffusion-based methods, are not used for registration per se but provide powerful semantic generation capabilities for repairing holes and artifacts produced after registration. Motivated by these observations, we propose a two-stage framework that combines geometric mapping with image inpainting. In the first stage, leveraging the precision of geometric vision, dense mapping from IR/depth images to the RGB space and the corresponding visible masks are generated using camera intrinsics, extrinsics, and depth information. In the second stage, conditioned on the target RGB image content, high-quality inpainting is performed in masked regions by combining a cross-modal attention mechanism with an advanced diffusion-based resampling strategy. This paradigm organically integrates the interpretability and geometric precision of parametric registration with the flexibility and content completeness of generative models, aiming to produce cross-modal images that are geometrically aligned and visually complete.

2.2 Driver Behavior Datasets for Driving Scenarios

Over the past decade, a variety of driver behavior datasets have been released, providing important support for driver state monitoring and behavior analysis research. However, as

shown in Table 1, most existing datasets focus on behavior classification tasks, whereas datasets specifically designed for cross-modal object detection with pixel-level registration remain extremely scarce.

Early studies mainly focused on single-modality RGB images. For example, distracted driving datasets such as SF-DDD [22], AUC-DD [23], and AVDM [24] provide a large number of daytime driving images. However, their annotations are image-level class labels, which are primarily suitable for behavior classification, and they generally lack nighttime scenarios, thus limiting algorithmic applicability in all-weather environments.

Subsequent studies began to introduce IR modalities. Both the NTHU-DDD dataset [25] and the 100-Driver dataset [26] contain RGB and IR data and cover both daytime and nighttime scenarios. Meanwhile, some researchers have also constructed datasets dominated by RGB and depth data, such as DAD [27]. With the development of sensor technology, tri-modal datasets containing RGB, IR, and depth modalities have gradually emerged, such as DriverMHG [28], Drive&Act [29], DMD [30], and MSHMD [31]. These datasets substantially enrich the categories of driving behaviors, reaching as many as 93 classes, and provide longer video recordings. TICaM [32] is an important exception, as it provides multi-task annotations including bounding boxes, instance segmentation masks, and 2D keypoints for RGB, IR, and depth data; however, its scenes are limited to daytime conditions.

Overall, existing driver behavior datasets still exhibit three major limitations when applied to driver behavior detection tasks. First, illumination and scene diversity are insufficient. Most datasets are collected under daytime or uniformly illuminated conditions and provide limited coverage of complex real-world driving environments such as nighttime, backlighting, strong reflections, and drastic lighting changes, thereby restricting model generalization in real applications. Second, there is a mismatch between task settings and annotation forms. Most datasets, such as NTHU-DDD, Drive&Act, DMD, and MSHMD, provide image-level labels or video clip labels, which are mainly suitable for classification rather than object detection. Their category systems are also organized around coarse behaviors such as “making a phone call,” rather than basic objects such as “hand” or “mobile phone,” making it difficult for models to learn fine-grained objects and their interactions. Although TICaM provides object detection annotations, such annotations are only available for the IR and depth modalities, while the RGB data are still used for classification tasks. More importantly, existing datasets generally lack cross-modal pixel-level registration.

Even when multiple modalities are included, the data are usually captured by independent, uncalibrated cameras and only roughly synchronized at the frame level. Without accurate camera intrinsics, extrinsics, and distortion models, precise pixel-level alignment cannot be achieved. As a result, prior studies are largely limited to feature-level or decision-level fusion strategies and are unable to investigate or evaluate the performance gains brought by pixel-level fusion.

To solve the above problems, this paper constructs a new multimodal driver behavior dataset, CMMID. This dataset is synchronously acquired in an intelligent cockpit environment and contains RGB, IR, and depth data with different capture periods, illumination conditions, target scales, pixel-level registration, and highly consistent annotations. The construction of this dataset is intended to fill the gap in existing research and to provide a solid data foundation for the development and evaluation of high-performance cross-modal pixel-level fusion detection algorithms.

3 Cross-Modal Registration Method

In the multimodal data preprocessing pipeline, cross-modal registration is a crucial step for subsequent fusion and analysis. Since registration quality directly affects the performance of downstream multimodal detection and behavior understanding tasks, this work develops a two-stage registration method centered on this key problem, as illustrated in Fig. 1. The method combines geometric mapping with mask-guided inpainting. In the first stage, strict pixel-level spatial correspondences are established through geometric mapping, and hole regions caused by occlusion or field-of-view differences are identified. In the second stage, a diffusion-based generative model integrating cross-modal guidance and frequency-domain enhancement is employed to repair these hole regions with semantically consistent and structurally complete content. Compared with traditional methods that rely solely on geometric mapping, the proposed method preserves high-precision alignment while producing visually complete registered images, thereby providing higher-quality inputs for downstream multimodal object detection and behavior understanding tasks.

3.1 Cross-Modal Geometric Parameterization and Hole Localization

The core objective of this stage is to establish an accurate geometric mapping from the IR image viewpoint to the RGB image viewpoint. Using pre-calibrated camera parameters, we formulate the registration problem as a deterministic three-dimensional coordinate transformation and reprojection process from the source image to the target image. Specifically,

Table 1 Comparison of representative driver behavior datasets

Dataset	Modality	Resolution	Scale	Annotation	Categories	Time	Registered	Year
SF-DDD	RGB	640 × 480	22,424 images	Image-level labels	10 driver behaviors	Daytime	No	2016
AUC-DD	RGB	1080 × 1920	11,678 images	Image-level labels	10 driver behaviors	Daytime	No	2017
AVDM	RGB	640 × 480	335k images	Image-level labels	8 driver behaviors	Daytime	No	2024
NTHU-DDD	RGB	1280 × 720	360 + 90 videos	Video-clip labels	8 driver behaviors	Day&night	No	2017
	IR	640 × 480						
100-Driver	RGB	1920 × 1080	470,208 images	Image-level labels	22 driver behaviors	Day&night	No	2023
	IR	—						
DAD	RGB	—	495 videos	Video-clip labels	11 driver behaviors	Daytime	No	2024
	DEPTH	—						
DriverMHG	RGB	320 × 240	5,250 videos	Video-clip labels	5 micro-hand gestures	Day&night	No	2020
	IR	320 × 240						
Drive&Act	DEPTH	320 × 240	About 12 h video	Video-clip labels	83 fine-grained behaviors	Day&night	No	2019
	RGB	950 × 540						
DMD	IR	512 × 424	41 h video	Video-clip labels	93 monitoring classes	Day&night	No	2020
	DEPTH	512 × 424						
TICaM	RGB	1920 × 1080	6,668 real pairs	Detection	19 cabin classes	Daytime	No	2021
	IR	1280 × 720						
MSHMD	DEPTH	512 × 512	3,306 synthetic pairs	Segmentation	19 cabin classes	Daytime	No	2021
	RGB	512 × 512						
CMMID (ours)	RGB	—	118k RGB images	Keypoints	10 driver behaviors	Day&night	No	2026
	IR	—						
CMMID (ours)	RGB	—	2,149 videos	Image-level labels	10 driver behaviors	Day&night	No	2026
	DEPTH	—						
CMMID (ours)	RGB	1280 × 720	139 videos	Detection labels	7 object classes	Day&night	Yes	2026
	IR	640 × 576						
CMMID (ours)	DEPTH	640 × 576	17,343 image groups	Detection labels	7 object classes	Day&night	Yes	2026
	DEPTH	640 × 576						

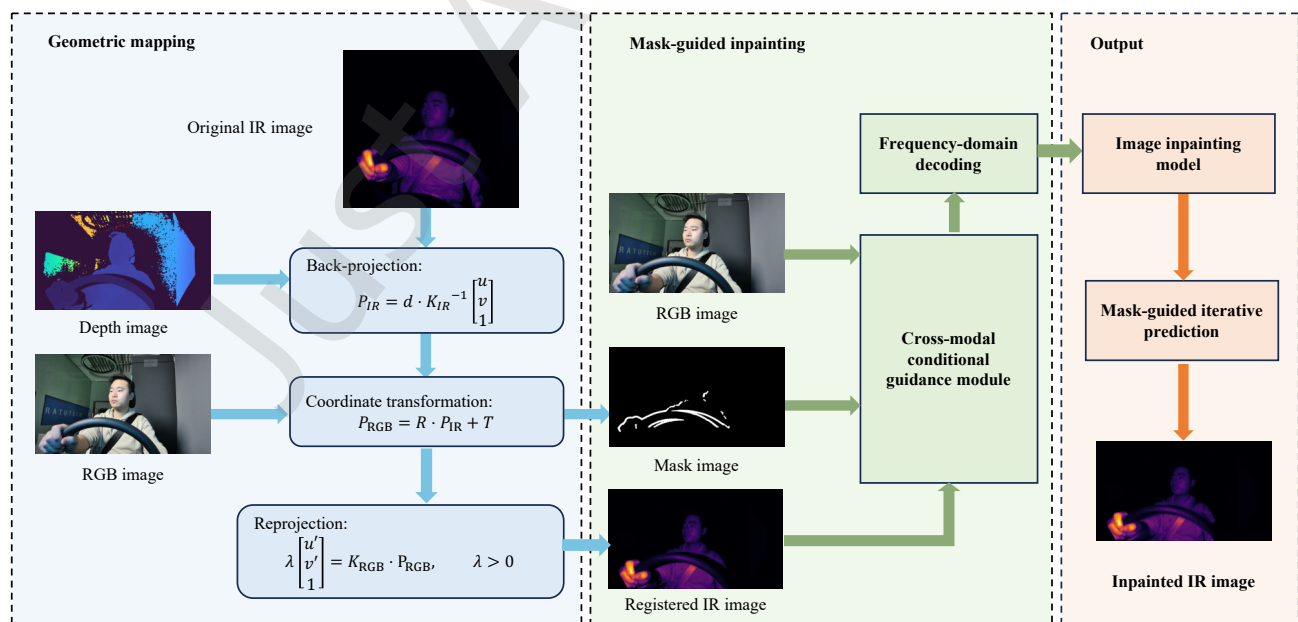


Fig. 1 Overall workflow of the proposed two-stage cross-modal registration method, read from left to right. In Stage 1, geometric mapping takes the original IR image and depth image and maps them to the RGB viewpoint through back-projection, coordinate transformation, and reprojection, producing a registered IR image together with a hole mask. In Stage 2, mask-guided inpainting takes the RGB image, the hole mask, and the registered IR image as inputs; cross-modal conditional guidance and frequency-domain decoding are then used to restore the missing regions. The final output is an inpainted IR image that is pixel-level aligned with the RGB image.

for any pixel in the IR image and its corresponding valid depth value, we first back-project it to a 3D point in the IR camera coordinate system using the inverse of the camera intrinsic matrix:

$$\mathbf{P}_{\text{IR}} = d \cdot \mathbf{K}_{\text{IR}}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (1)$$

where $\mathbf{P}_{\text{IR}} \in \mathbb{R}^3$ denotes the 3D point in the IR camera coordinate system, d denotes the valid depth value at pixel (u, v) , $\mathbf{K}_{\text{IR}}$ denotes the intrinsic matrix of the IR camera, and $[u, v, 1]^\top$ denotes the homogeneous pixel coordinate in the IR image plane.

Then, using the calibrated inter-camera extrinsic parameters, namely the rotation matrix and translation vector, the 3D point is rigidly transformed from the IR coordinate system to the RGB camera coordinate system, yielding its corresponding point in the target coordinate system:

$$\mathbf{P}_{\text{RGB}} = \mathbf{R} \cdot \mathbf{P}_{\text{IR}} + \mathbf{T} \quad (2)$$

where $\mathbf{P}_{\text{RGB}} \in \mathbb{R}^3$ denotes the transformed 3D point in the RGB camera coordinate system, $\mathbf{R} \in SO(3)$ denotes the rotation matrix from the IR camera to the RGB camera, and $\mathbf{T} \in \mathbb{R}^3$ denotes the corresponding translation vector.

Finally, the intrinsic matrix of the RGB camera is used to project the 3D point onto the image plane. After normalization, the precise corresponding pixel coordinates in the RGB image are obtained:

$$\lambda \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \mathbf{K}_{\text{RGB}} \cdot \mathbf{P}_{\text{RGB}}, \quad \lambda > 0 \quad (3)$$

where (u', v') denotes the corresponding pixel coordinate in the RGB image, λ denotes the positive projective scale factor, and $\mathbf{K}_{\text{RGB}}$ denotes the intrinsic matrix of the RGB camera.

By applying the above transformation to all pixels, we construct a dense mapping field from the IR viewpoint to the RGB viewpoint. To generate the aligned IR image while minimizing resampling holes, we adopt an inverse warping strategy with bilinear interpolation. If a depth value is invalid or the computed projection falls outside the image boundary, the corresponding location in the aligned IR image is regarded as a registration hole. Based on this criterion, a binary hole mask is generated, in which pixels with value 1 denote holes. This mask and its complement, corresponding to valid regions, serve as key conditional inputs to the restoration model in the second stage.

3.2 Multimodal Conditional-Guided Image Inpainting

To fill the hole regions produced in the first stage, this stage employs a generative model for high-quality content completion. In this preprocessing pipeline, the image inpainting component integrates the prediction strategy of diffusion models, a frequency-enhanced network architecture, and a multimodal conditional guidance mechanism. The overall framework is shown in Fig. 2(a). The backbone of the model is a classical U-Net architecture, which effectively combines deep semantic information and shallow texture details through an encoder-decoder structure with skip connections. To enable joint spatial-frequency modeling, enlarge the receptive field, and strengthen global structural perception over large missing regions, we further modify the standard U-Net by replacing the first two convolutional modules in the decoder with Fast Fourier Convolution (FFC) modules.

The encoder contains four stages, each composed of multiple residual blocks. After each stage, a cross-modal conditional guidance module is introduced, in which the RGB image serves as the guiding modality and is fused with IR features to provide structural and semantic completion cues at an early stage, thereby improving the overall inpainting quality. The multi-level features are progressively downsampled by the encoder and then fed into the bottleneck module, where deeper semantic information is further aggregated and the key prediction step of the diffusion model is executed. The features are subsequently sent to the decoder, whose structure is symmetric to that of the encoder. Through multi-stage upsampling and skip connections, the spatial resolution is gradually recovered, and a structurally complete and semantically clear IR image is finally produced.

To effectively inject structural and semantic information from the RGB image into the IR image inpainting process, we design a cross-modal conditional guidance module, as shown in Fig. 2(b). In this module, the RGB and IR images are first processed by a shared encoder to extract multi-scale features, which are aligned to a 64-dimensional channel space. Then, after each of the four encoder stages, guided fusion is performed using multi-head scaled dot-product attention, where IR features act as queries and RGB features act as keys and values. A four-head attention design is adopted, with 16 dimensions per head. The resulting outputs are concatenated into 64 dimensions, concatenated with the original IR features along the channel dimension to form a 128-dimensional representation, and then integrated by a convolution layer followed by ReLU. The fused features are further passed to the corresponding decoder layer through skip connections. In this

way, the model can adaptively capture structural priors such as contours and boundaries, as well as high-level semantic information, from the RGB modality and incorporate them into the generation of IR features. This design allows the IR branch to receive accurate structural and semantic guidance at an early stage and significantly improves the restoration quality.

To further enhance global consistency modeling for large hole regions, we introduce FFC modules into the first two decoding layers of the U-Net, thus enabling joint spatial–frequency representation. FFC divides the input channels into a local branch and a global branch. The local branch uses conventional convolution to process local information, whereas the global branch uses the real-valued FFT (Real FFT) to model global context in the frequency domain. Real FFT is applicable only to real-valued signals, and inverse Real FFT guarantees that the output remains real-valued. Compared with the standard complex FFT, Real FFT uses only half of the spectrum. The detailed process is as follows:

First, Real FFT2d is applied to the input tensor, and the real and imaginary parts are concatenated along the channel dimension:

$$\mathbb{R}^{H \times W \times C} \rightarrow \mathbb{C}^{H \times \frac{W}{2} \times C} \quad (4)$$

$$\mathbb{C}^{H \times \frac{W}{2} \times C} \rightarrow \mathbb{R}^{H \times \frac{W}{2} \times 2C} \quad (5)$$

where H and W denote the spatial height and width of the feature map, C denotes the number of channels, $\mathbb{R}$ and $\mathbb{C}$ denote real-valued and complex-valued feature spaces, respectively, and the second mapping denotes concatenation of the real and imaginary parts along the channel dimension.

Next, a convolution block is applied in the frequency domain, after which the real and imaginary parts are split along the channel dimension and reassembled into a complex representation:

$$\mathbb{R}^{H \times \frac{W}{2} \times 2C} \rightarrow \mathbb{R}^{H \times \frac{W}{2} \times C} \quad (6)$$

$$\mathbb{R}^{H \times \frac{W}{2} \times 2C} \rightarrow \mathbb{C}^{H \times \frac{W}{2} \times C} \quad (7)$$

where the first mapping denotes channel mixing in the frequency domain after convolution, and the second mapping denotes reassembling the split real and imaginary components into a complex-valued spectrum.

The result is then transformed back to the spatial domain through the inverse transform to restore the spatial structure:

$$\mathbb{C}^{H \times \frac{W}{2} \times C} \rightarrow \mathbb{R}^{H \times W \times C} \quad (8)$$

where the inverse Real FFT maps the complex frequency representation back to a real-valued spatial feature map with the original resolution.

Finally, the outputs of the local and global branches are fused together. As a fully differentiable plug-and-play module, FFC is crucial for high-resolution image inpainting because its global receptive field helps preserve overall consistency while improving computational efficiency.

To comprehensively optimize the inpainting result, we construct a multi-task composite loss function, as shown in Eq. (9). The loss for multimodal conditional-guided inpainting consists of a noise prediction loss, a perceptual loss, a hole loss, and a structure-preserving edge loss. Only the IR output is optimized; the RGB image participates only as a conditional feature and does not enter the target-side loss.

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{noise}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \mathcal{L}_{\text{hole}} + \lambda_4 \mathcal{L}_{\text{edge}} \quad (9)$$

where λ_1 – λ_4 are scalar weights used to balance denoising fidelity, perceptual consistency, hole reconstruction, and edge preservation, respectively. Among these terms, the noise prediction loss is the mean squared error (MSE) loss, as defined in Eqs. (10) and (11). This term is used to train the conditional U-Net to perform accurate denoising at any time step. Following the forward diffusion process, a noisy sample is generated from the reference ground-truth image according to the cumulative diffusion coefficient. The conditional network then predicts the noise term:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (10)$$

$$\mathcal{L}_{\text{noise}} = \mathbb{E}_{x_0, t, \epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2 \right] \quad (11)$$

where x_0 denotes the reference ground-truth IR image, x_t denotes the noisy sample at diffusion step t , $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ denotes the cumulative product coefficient of the diffusion schedule, ϵ denotes Gaussian noise, $\epsilon_\theta(\cdot)$ denotes the noise predicted by the conditional U-Net with parameters θ , and c denotes the cross-modal conditional feature extracted from RGB–IR guidance.

To preserve consistency with the reference image in terms of high-level semantics and texture appearance, we further introduce a perceptual loss based on a VGG19 network pretrained on ImageNet. Specifically, feature representations are extracted from the relu3_3 and relu4_2 layers, and the L2 distance between the generated image and the ground-truth image is computed to measure perceptual discrepancy:

$$\mathcal{L}_{\text{perc}} = \sum_{l \in \{\text{relu3.3}, \text{relu4.2}\}} \|\phi_l(\hat{\mathbf{x}}) - \phi_l(\mathbf{x}_0)\|_2^2 \quad (12)$$

where $\hat{\mathbf{x}}$ denotes the generated IR image, $\mathbf{x}_0$ denotes the ground-truth IR image, and $\phi_l(\cdot)$ denotes the feature extractor corresponding to layer l of the pretrained VGG19 network.

To reduce pixel-wise errors within hole regions, avoid brightness drift, and promote seamless transitions to known

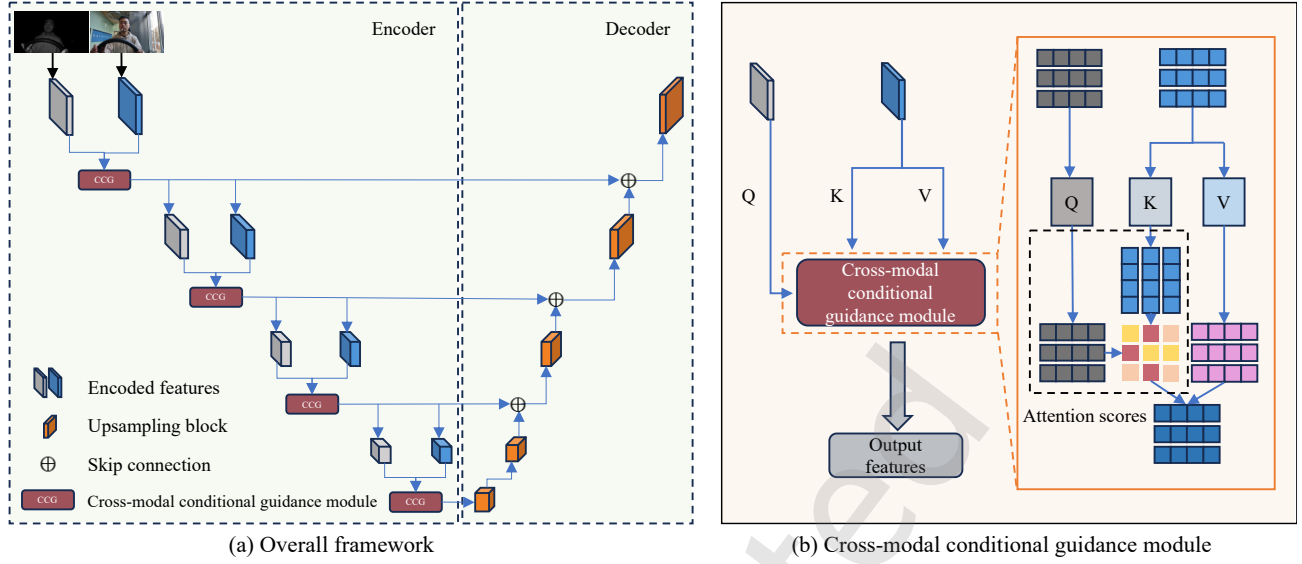


Fig. 2 Multimodal conditional-guided image inpainting framework: (a) the overall diffusion-based inpainting network; (b) the cross-modal conditional guidance module.

regions, we introduce a hole loss. This term computes the pixel-level L1 reconstruction error inside the hole area and normalizes it by the number of masked pixels:

$$\mathcal{L}_{\text{hole}} = \frac{1}{\max(1, \sum M)} \|M \odot (\hat{\mathbf{x}} - \mathbf{x}_0)\|_1 \quad (13)$$

where $M \in \{0, 1\}^{H \times W}$ denotes the hole mask, with 1 indicating missing pixels, and $\odot$ denotes element-wise multiplication. In addition, to maintain the continuity and sharpness of structural edges, a structure-preserving edge loss is introduced. This loss computes horizontal and vertical gradients using Sobel operators, extracts edge maps from the generated image and the ground-truth image, and imposes an L1 constraint within a boundary-band mask, as defined in Eqs. (14) and (15):

$$\mathbf{B} = \text{Dilate}(M, r) \setminus M, \quad r \in \{8, 16\} \text{ pixels} \quad (14)$$

$$\mathcal{L}_{\text{edge}} = \frac{1}{\max(1, \sum \mathbf{B})} \left(\|\mathbf{B} \odot (G_x(\hat{\mathbf{x}}) - G_x(\mathbf{x}_0))\|_1 + \|\mathbf{B} \odot (G_y(\hat{\mathbf{x}}) - G_y(\mathbf{x}_0))\|_1 \right) \quad (15)$$

where $\mathbf{B}$ denotes the boundary-band mask constructed by dilating M with radius r and removing the original hole region, while $G_x(\cdot)$ and $G_y(\cdot)$ denote the horizontal and vertical Sobel gradient operators, respectively.

Based on the reverse sampling process of the denoising diffusion probabilistic model (DDPM), this work performs progressive completion of hole regions through a mask-guided multi-round prediction strategy. The overall inference process is illustrated in Fig. 3. Let x^{in} denote the registered IR image used as input. A Gaussian noise sequence is first introduced

to construct the initial state, after which the standard DDPM reverse sampling process is adopted to iteratively denoise the sample from the initial noisy state to the final reconstructed output. After each iteration, a mask-guided fusion strategy is applied so that only the pixels in the hole region are updated while the known region remains unchanged. The operation is defined as follows:

$$\hat{x}_{t-1} = M \odot \hat{x}_{t-1}^{\text{pred}} + (1 - M) \odot x^{\text{in}} \quad (16)$$

where M denotes the same hole mask as in Eq. (13), $1 - M$ denotes its complementary valid-region mask, $\hat{x}_{t-1}^{\text{pred}}$ denotes the denoised prediction produced by the model at step $t - 1$, x^{in} denotes the original registered IR input, and $\hat{x}_{t-1}$ denotes the fused sample that is fed into the next iteration. This operation ensures that non-hole regions always preserve the original input content, effectively preventing the model from perturbing valid structures. Through the iterative loop of “prediction–mask fusion–iterative update,” the model ultimately produces results that are structurally continuous, semantically natural, and geometrically consistent with the original RGB image, making them directly applicable to subsequent multimodal object detection, behavior understanding, and multimodal fusion tasks.

4 Dataset Construction and Analysis

To address the problems of limited modality diversity, restricted illumination scenarios, and the lack of pixel-level registration in existing driver behavior datasets, we construct and analyze a new multimodal driver behavior dataset for cross-modal object detection tasks. This section describes

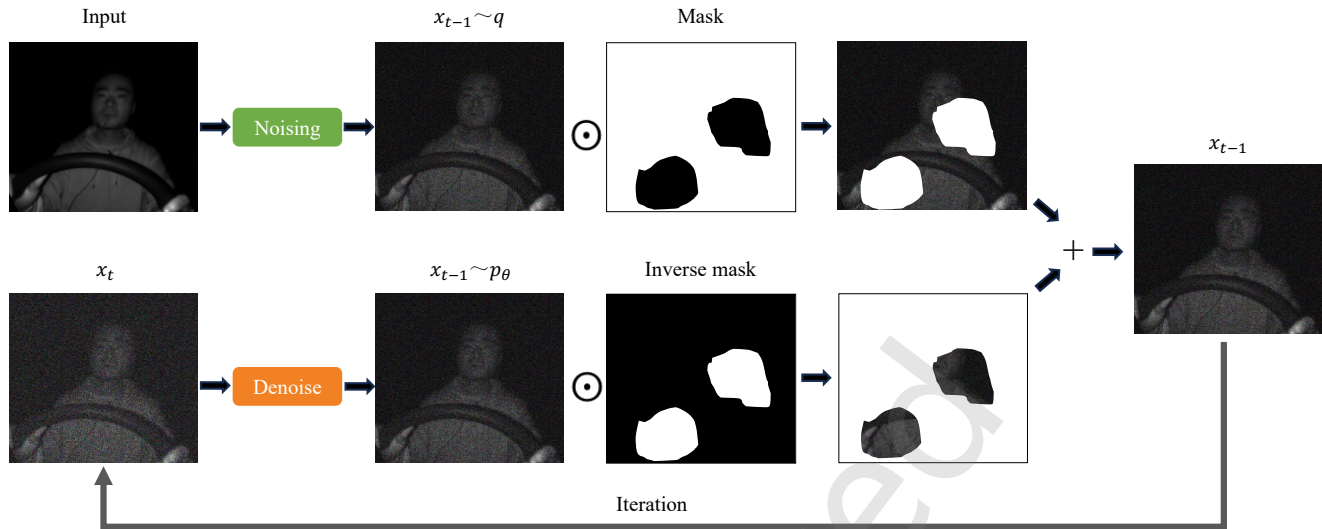


Fig. 3 Mask-guided iterative prediction process used during inference.

the complete dataset construction pipeline, including the data acquisition scheme, data preprocessing and annotation procedure, and an in-depth analysis of the statistical characteristics, diversity, and label quality of the dataset.

4.1 Data Acquisition Scheme

4.1.1 Experimental Platform

Driver monitoring systems often face a variety of challenging application scenarios, such as hazardous driver operations, non-ideal illumination, adverse weather, and day–night driving. These scenarios involve potential safety risks and place high demands on the safety and reliability of data collection. To obtain high-quality and diverse data in a controllable and safe environment, this study employs a 1:1 intelligent simulation cockpit developed by CITIC Simulation as the experimental platform and uses the Orbbec Femto-Bolt iToF 3D camera for data acquisition. The complete data acquisition setup is shown in Fig. 4. The intelligent simulation cockpit is equipped with a six-degree-of-freedom motion system that can realistically reproduce vehicle acceleration and deceleration forces, cornering centrifugal effects, and road bumps, thereby recreating real driving states in a safe environment and significantly improving the dynamic realism and scenario diversity of the collected data. The main specifications of the Femto-Bolt camera are listed in Table 2.

The camera was fixed approximately 25 cm behind the steering wheel on the driver side and oriented toward the driver with a downward tilt angle of approximately 15°, so as to capture a complete view of the face and upper body while avoiding occlusion of the instrument cluster and head-up display. The camera can synchronously output RGB,

IR, and depth streams from a single hardware device and provides hardware-level timestamp synchronization, enabling simultaneous acquisition of aligned tri-modal frames and thus laying the foundation for subsequent cross-modal alignment and registration. In addition, the ToF depth measurement range of the Femto-Bolt camera is 0.5–3.86 m, which covers the typical working distance from the driver’s upper body to the steering wheel and dashboard and can accurately capture three-dimensional structural information of the hands and body posture.

4.1.2 Scenario Design

To obtain representative driving samples, the data collection process was designed hierarchically with illumination conditions as the main organizing axis. Specifically, four illumination levels were defined: daytime, nighttime without obvious supplementary lighting, nighttime with lights (simulating streetlights, cabin dome lights, or strong instrument illumination), and excessive lighting / strong reflection (simulating direct strong light, specular highlights, or local overexposure caused by bright screens). The definitions and examples of these four illumination conditions are listed in Table 3.

Taking smoking, a hazardous driving behavior, as an example, the driver in each illumination condition followed a unified experimental script consisting of vehicle start-up → cruising → smoking-related operations (including lighting, inhaling, and flicking ash) → parking, so as to ensure consistency in the temporal structure and behavioral pattern of the samples. Throughout the collection process, the camera synchronously recorded multimodal information such as facial expressions, gaze changes, and hand operations, thereby



Fig. 4 Intelligent cockpit data acquisition platform equipped with the Orbbec Femto-Bolt camera and representative data samples.

Table 2 Main specifications of the Femto-Bolt camera

Device	Modality	Resolution @ 30 fps	Band
Orbbec Femto-Bolt	RGB	1280 × 720 @ 30 fps	RGB
	IR	640 × 576 @ 30 fps	850 nm
	Depth	640 × 576 @ 30 fps	850 nm iToF

Table 3 Illumination categories used during data collection

Illumination level	Definition	Typical source
Daytime	High average brightness without large saturated highlight regions.	Natural daylight
Nighttime (dark)	Low average brightness, with only weak interference from screens or road lights.	No streetlights; cabin lights off
Nighttime (lit)	Streetlights or cabin/dashboard lighting improve the visibility of the driver's face and hands.	Lit roads or garage exits
Excessive lighting	Large bright regions with obvious local saturation, glare, or specular reflection.	Oncoming headlights; reflections from screens or trim

providing high-quality data support for subsequent temporal behavior analysis and cross-modal feature alignment.

4.1.3 Participants

A total of 16 volunteers participated in this study, including 14 males and 2 females, with ages ranging from 20 to 35 years. All participants were non-professional drivers holding valid driving licenses, with actual driving experience ranging from 1 to 8 years. During the experiment, each participant followed unified instructions from the data collection staff and sequentially completed the simulated driving and smoking operations under the aforementioned illumination scenarios. The recording duration for each scenario was approximately 90 seconds, ensuring coverage of complete driving and smoking action sequences. In total, 139 raw synchronized video sequences in ".mkv" format were collected, with a total volume of approximately 282 GB. Each raw video contains a complete driving process together with the corresponding

hazardous driving behavior segment, providing a high-quality data basis for subsequent frame-level annotation, feature extraction, and model training.

4.2 Data Preprocessing and Annotation

After encapsulation and storage, the raw video data must undergo systematic preprocessing and registration before image frames and annotation files suitable for model training can be obtained. The overall preprocessing pipeline mainly includes demultiplexing and sample screening.

First, this study uses C++ together with OpenCV and the Azure Kinect Sensor SDK to decode the raw video files frame by frame, with an average processing speed of approximately 0.15 s/frame. During decoding, the program names the image sequences according to the recording timestamps and separates the synchronously output RGB, IR, and depth images, thereby laying the foundation for subsequent frame-wise pro-

cessing. A total of 518,217 groups of raw frame sequences were obtained. Considering dataset diversity and annotation cost, a preliminary frame selection strategy of sampling one frame every ten frames was adopted along the temporal dimension. Subsequently, samples with high redundancy, overly monotonous scenes, or severe blur were manually removed. After the two-stage screening process, approximately 90% of highly redundant frames and some frames with limited value for model training were discarded, ensuring that the retained images were suitable for object detection model training. Ultimately, 17,343 groups of high-quality raw frame samples were obtained (52,029 images in total across the three modalities), which can be directly used for subsequent multimodal object detection and driver behavior understanding tasks.

Because the three modalities originate from different lenses with slightly different physical positions, fine-grained registration was performed on the preprocessed data according to the method described in Section 3 to ensure spatial consistency among modalities at the same timestamp. Temporal synchronization had already been achieved in the demultiplexing stage through the unified timestamp-based naming mechanism. For spatial registration, the RGB, IR, and depth imaging chains were first calibrated separately to obtain their respective intrinsic and distortion parameters, after which all images were undistorted and normalized under a unified coordinate convention. Then, cross-modal extrinsic parameters were estimated using multi-view calibration-board images, yielding the rigid transformation from the IR/depth cameras to the RGB camera coordinate system. During registration, the RGB camera coordinate system was taken as the target. For the depth image, each valid pixel was first back-projected to 3D using the intrinsic parameters, transformed into the RGB coordinate system by the extrinsic parameters, and then projected onto the image plane, thereby realizing dense depth-to-visible mapping. For the IR image, the same extrinsic framework was used to perform the geometric mapping from IR to RGB space. To address many-to-one mappings and holes that may arise during projection, this study adopted inverse mapping with bilinear interpolation and used a depth-priority Z-buffer strategy to handle occlusion conflicts. Through this process, the three modalities were finally unified into the RGB camera coordinate system and resolution, forming pixel-level aligned multimodal samples and providing an accurate data basis for joint model training and multimodal fusion analysis.

During the annotation stage, we organized a team of 12 annotators to manually label all 17,343 RGB frames. Thanks to the strict temporal synchronization and cross-modal registration described above, the annotations created on the RGB

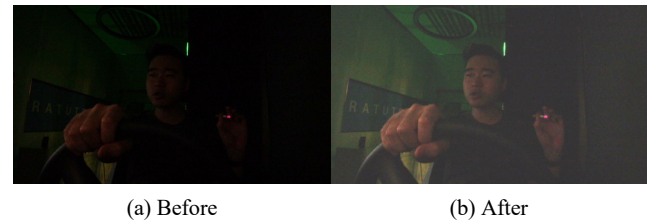


Fig. 5 Examples before and after brightness enhancement used during annotation.

modality could be synchronously transferred to the corresponding IR and depth images, thus maintaining consistency across the three modalities. Although the annotations were primarily performed on RGB images, annotators could simultaneously inspect registered pseudo-color IR/depth reference images in order to improve decision accuracy in scenes where foreground and background intensities were similar or object boundaries were ambiguous. For nighttime samples with an average brightness below 40, we introduced linear brightness enhancement during the annotation stage and increased the brightness to approximately the 60–80 range, generating “brightened” images for annotation to improve visibility and reduce missed or incorrect labels caused by insufficient illumination. It should be emphasized that this enhancement was used only for annotation and did not alter the image geometry or coordinates. Fig. 5 shows a comparison before and after low-light enhancement. It can be seen that the originally dark image becomes visually distinguishable after enhancement, making it easier for annotators to label the targets accurately. Through the above preprocessing and annotation procedure, a high-quality, temporally synchronized, and pixel-level registered multimodal driver behavior dataset was finally constructed.

4.3 Dataset Statistics and Illumination Characteristic Analysis

Table 4 summarizes the overall statistics of the dataset constructed in this study. The dataset contains a total of 52,029 valid images (17,343 per modality), collected from 16 drivers (14 males and 2 females, aged 20–35 years), with annotations for seven behavior-related categories and a total of 85,787 labels. The RGB image resolution is 1280×720 , while the raw IR and depth image resolutions are 640×576 , all captured at 30 fps with strict synchronization across the three modalities. After registration, the IR and depth images are resampled to 1280×720 to achieve pixel-level alignment across modalities. These data expand the scale of existing driver behavior datasets from multiple dimensions, including inter-driver variation, category diversity, and illumination

diversity.

To meet the needs of multimodal hazardous behavior detection, we designed an annotation system containing seven fine-grained categories. These categories are head, hand, safety belt, phone, bottle, cigarette, and steering wheel, all of which are closely related to driver behavior analysis and hazardous driving detection. Table 5 summarizes the category-wise label counts together with their image-level occurrence ratios, where the ratio denotes the percentage of annotated images containing at least one instance of the corresponding category. As shown in the statistics, the dataset covers both common driver/body regions and hazardous objects, providing a diverse annotation basis for multimodal detection and behavior analysis.

In addition, to analyze the characteristics of different modalities under diverse illumination conditions, Fig. 6 presents typical tri-modal RGB/IR/depth image samples. In the daytime scene with sufficient illumination (Fig. 6(a)), the RGB image clearly presents driver details. In the nighttime scene without lights (Fig. 6(b)), the RGB image is almost completely dark and difficult to interpret, whereas the corresponding IR image still clearly shows facial and hand contours, and the depth image provides structural contour information of the driver's posture, demonstrating the complementary value of IR and depth modalities under low-light conditions. In the nighttime scene with lights (Fig. 6(c)), RGB images still preserve recognizable driver details. In the excessive-lighting nighttime scene (Fig. 6(d)), which simulates direct oncoming vehicle lights entering the cabin, the responses of different modalities to strong illumination interference can be observed. Specifically, the RGB image exhibits clear overexposure and loss of local details, whereas the IR and depth images remain relatively stable and robust against interference. These observations show that under extremely low or highly variable illumination, a single RGB modality alone is insufficient for timely and accurate detection of driver-related objects and cues, whereas integrating IR and depth information can significantly improve detection robustness and reliability.

4.4 Annotation Quality Evaluation

To ensure high annotation quality and reliability, the dataset was annotated independently twice, and Cohen's kappa coefficient was used to quantitatively evaluate annotation consistency. The results show that the Cohen's κ coefficient between the two rounds of independent annotation reached 0.9974, which corresponds to an "almost perfect agreement" level. This indicates extremely high consistency among annotators in both category recognition and bounding-box localization,

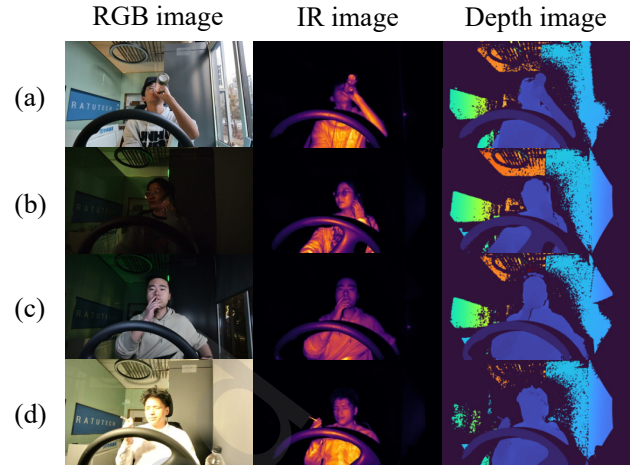


Fig. 6 Representative multimodal samples under different illumination conditions: (a) daytime, (b) nighttime without lights, (c) nighttime with lights, and (d) excessive lighting / strong reflection.

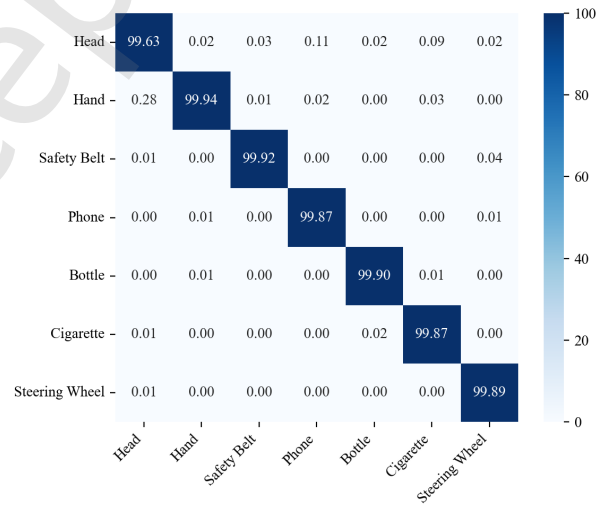


Fig. 7 Normalized confusion matrix between the two rounds of independent annotation.

demonstrating the reliability of the dataset labels. Fig. 7 presents the confusion matrix between the two annotation rounds to visualize the distribution of discrepancies. As can be seen, except for a slight confusion of about 0.28% between the head and hand categories, almost no significant deviations appear among the remaining categories. This result further verifies the accuracy and consistency of the annotation system and provides a high-quality annotation basis for subsequent model training.

Table 4 Summary statistics of the constructed CMMID dataset

Item	Value	Notes
Participants	16 drivers	14 males and 2 females; age range 20–35 years
Modalities	RGB, IR, and depth	Synchronously captured by a single Orbbec Femto-Bolt device
Image groups	17,343	Each group contains one synchronized RGB–IR–depth triplet
Total images	52,029	17,343 images were retained for each modality after screening
Categories	7 object classes	Detailed category definitions are listed in Table 5
Annotations	85,787 labels	Labels were annotated on RGB images and then transferred to registered IR/depth images
Acquisition rate	30 fps	Strict temporal synchronization was maintained across all three modalities
Resolution	RGB 1280×720	Registered IR images were resampled to 1280×720 for pixel-level alignment.
	Raw IR 640×576	Registered depth images were resampled to 1280×720 for pixel-level alignment.
	Raw depth 640×576	

5 Experiments and Results Analysis

5.1 Experimental Setup

5.1.1 Dataset

The experiments in this study are conducted on the self-constructed CMMID intelligent-cockpit multimodal image dataset. Derived from real driving behaviors, this dataset covers diverse illumination and pose variations and therefore exhibits high complexity and representativeness. Following an 8:1:1 split, the CMMID dataset is divided into training, validation, and test sets. The training set contains 13,875 groups of RGB–IR–depth images and 69,329 instance annotations, the validation set contains 1,734 groups of RGB–IR–depth images and 8,577 instance annotations, and the test set contains 1,734 groups of RGB–IR–depth images and 7,881 instance annotations.

5.1.2 Experimental Environment and Parameter Settings

All experiments in this study were conducted on a high-performance server equipped with two AMD EPYC 7543 processors (32 cores / 128 threads each), 1.0 TB of system

memory, and eight NVIDIA A800-SXM4-80GB GPUs. The server ran Ubuntu 20.04.6 LTS, with GPU driver version 550.54.15, CUDA 11.8, PyTorch 2.1.0, and Python 3.10.14. All training and evaluation experiments for the object detection models were carried out under Ultralytics version 8.3.191. The entire study was performed under a consistent hardware and software environment to ensure experimental reproducibility. For fair comparison and reproducibility, the same training protocol was adopted for YOLOv5 and YOLOv8 under all four modality settings, namely RGB, RGB + IR, RGB + depth, and RGB + IR + depth. Specifically, all models were trained on the training split for 300 epochs and validated on the validation split, while the final results were reported on the test split. The input resolution was fixed at 640×640 pixels, and the batch size was set to 32. Stochastic gradient descent (SGD) was used as the optimizer, with an initial learning rate of 0.1 and a weight decay of 0.0005. During training, Mosaic augmentation was enabled in the main training stage and disabled in the last 50 epochs, while random perturbation and random flipping were uniformly applied to aligned RGB, IR, and depth images to preserve cross-modal correspondence. Unless otherwise stated, no additional modality-specific hyperparameter adjustment was introduced, and the compared models differed only in the input modality combination and backbone architecture.

5.2 Evaluation Metrics

To systematically evaluate the robustness and applicability of the algorithm under non-ideal illumination conditions, we use Precision (P), Recall (R), Average Precision (AP) for each category, and mean Average Precision (mAP) to evaluate detection performance. Precision reflects discriminative accuracy: a higher value indicates more accurate target judgment and fewer false positives. Recall reflects detection capability:

Table 5 Category-wise label distribution in CMMID

Category	Label count	Ratio
Head	17,928	100%
Hand	26,866	100%
Safety belt	7,414	42.75%
Phone	5,471	31.55%
Bottle	4,132	23.83%
Cigarette	6,633	38.25%
Steering wheel	17,343	100%

Note: Ratio refers to the proportion of images containing at least one instance of the corresponding category among the total number of images. Therefore, the reported value of 100% for head, hand, and steering wheel means that these categories are present throughout the full annotated image set considered in this table. Since more than one instance may appear in a single image for the head and hand categories, their label counts may exceed the total number of images.

a higher value indicates stronger detection ability and fewer missed targets. A higher AP for a given category indicates better detection performance on that category, whereas a higher mAP indicates stronger overall detection performance and robustness under complex illumination. The corresponding formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (17)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (18)$$

where TP , FP , and FN denote the numbers of true positives, false positives, and false negatives, respectively. When the IoU between a predicted box and a ground-truth box is greater than or equal to τ , the prediction is regarded as a correct match, and each ground-truth instance can be matched to at most one prediction. AP is defined as the area under the Precision–Recall curve obtained by varying the confidence threshold for a given category and IoU threshold. mAP is then computed as the arithmetic mean of AP over all categories. In this paper, we report two commonly used evaluation protocols: mAP@0.5 (mAP50), which averages AP over all categories at a fixed IoU threshold of 0.50 and reflects overall detection performance under a relatively loose matching criterion; and mAP@0.5:0.95 (mAP50–95), which follows the COCO protocol and computes AP at multiple IoU thresholds in $\{0.50, 0.55, \dots, 0.95\}$ with a step size of 0.05, averages over IoU thresholds first, and then averages over categories. The latter is a stricter metric that jointly measures both classification correctness and localization precision.

5.3 Comparative Experiments

5.3.1 Qualitative Comparison of Registration Method

We perform a qualitative comparison between the traditional registration method and the proposed CMMIR, as shown in Fig. 8. The traditional registration method produces noticeable hole artifacts and boundary discontinuities in the aligned results. This phenomenon mainly arises from viewpoint differences, occlusion, and inconsistencies in imaging coverage across modalities when infrared/depth images are mapped to the RGB viewpoint. In particular, holes may also appear around key behavior-related objects, such as hands and cigarettes, leading to missing visual cues, fragmented local structures, and incomplete object boundaries. These defects reduce the visual integrity of target regions and may further impair downstream multimodal detection by lowering detection confidence and localization accuracy. In contrast, the results produced by CMMIR exhibit better boundary continuity and stronger semantic consistency. By effectively

restoring plausible content in hole regions, CMMIR improves the overall visual quality of target areas and provides more reliable inputs for subsequent multimodal detection tasks.

5.3.2 Comparative Experiments Based on YOLOv5

To verify the effectiveness of the multimodal image data in the constructed dataset for multimodal object detection for driver behavior understanding, we first conducted comparative experiments using YOLOv5 as the baseline model. The experiments were performed under four modality combinations: RGB, RGB + IR, RGB + depth, and RGB + IR + depth. All experiments were conducted on NVIDIA A800 GPUs and evaluated under the same protocol on the test set. The results are shown in Table 6, where the values in parentheses indicate the absolute improvement over the RGB baseline in percentage points.

Compared with single-modality object detection models, multimodal models show significant advantages in overall detection accuracy. In terms of precision (P) and recall (R), the RGB-only model achieves 57.1% and 33.9%, respectively. After introducing the depth modality, P and R increase by 0.4 and 15.3 percentage points, respectively. After introducing the IR modality, the gains are even more pronounced, with P and R improving by 32.7 and 28.3 percentage points, respectively. When IR and depth information are jointly fused in the tri-modal configuration (RGB + IR + depth), P and R reach 90.9% and 84.2%, representing improvements of 33.8 and 50.3 percentage points over the RGB-only baseline and achieving the best performance among all settings. These results suggest that depth cues help reduce missed detections, while infrared information is particularly beneficial under challenging illumination conditions. Compared with the single-modality R baseline, the RGB + IR configuration improves mAP@0.5 and mAP@0.5:0.95 by 32.5 and 25.1 percentage points, respectively, while the RGB + depth configuration improves these two metrics by 11.3 and 10.0 percentage points, respectively. The tri-modal RGB + IR + depth configuration achieves the largest gains, improving mAP@0.5 and mAP@0.5:0.95 by 50.6 and 36.8 percentage points, respectively, which further validates the effectiveness of multi-source fusion.

The results indicate that IR images possess stronger illumination invariance and more discriminative skin/material reflectance characteristics, which significantly improve the visibility of small or low-contrast targets such as hands, phones, cigarettes, and safety belts under low-light and backlit conditions. Depth images, by providing stable geometric boundaries and foreground–background separation cues, can effectively reduce false detections. Together, these modalities

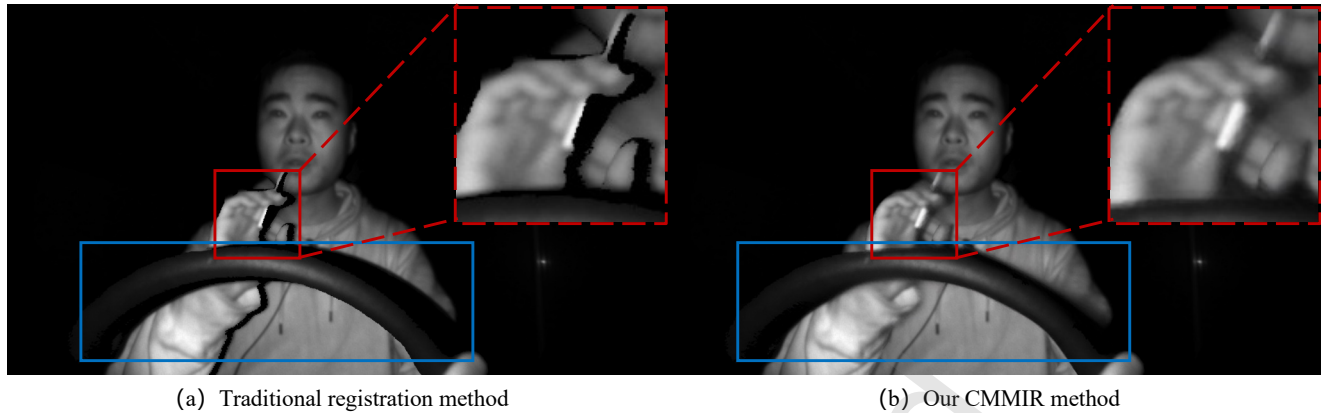


Fig. 8 Qualitative comparison between the traditional registration method and CMMIR. The traditional method produces noticeable hole artifacts around key targets, which disrupt local structures and weaken visual cues, whereas CMMIR restores structurally continuous and semantically consistent results that provide more reliable inputs for downstream detection.

Table 6 Detection results of different modality combinations based on YOLOv5

Backbone	Modality	AP (%)							P (%)	R (%)	mAP50 (%)	mAP50–95 (%)
		Head	Hand	Safety belt	Phone	Bottle	Cigarette	Steering wheel				
YOLOv5	RGB	90.7	26.9	21.5	5.4	12.3	11.5	98.3	57.1	33.9	38.1	26.3
YOLOv5	RGB + depth	96.9	56.5	27.4	20.6	22.4	22.4	99.7	57.5	49.2	49.4 (+11.3)	36.3 (+10.0)
YOLOv5	RGB + IR	99.5	74.7	35.7	63.0	64.0	57.8	99.7	89.8	62.1	70.6 (+32.5)	51.4 (+25.1)
YOLOv5	RGB + IR + depth	99.7	97.4	85.6	77.5	87.8	73.1	100.0	90.9	84.2	88.7 (+50.6)	63.1 (+36.8)

form a complementary relationship with the RGB modality at both the spectral and structural levels, jointly enhancing the discriminative ability and robustness of the model and yielding the largest overall performance gain. These results demonstrate that the multimodal image dataset constructed in this work can effectively improve multimodal object detection performance in intelligent cockpit scenarios.

5.3.3 Comparative Experiments Based on YOLOv8

To further verify the general applicability of the dataset across different architectures, we next used YOLOv8 as the baseline model. The experimental settings remained the same as in the previous subsection, with four modality combinations: RGB, RGB + IR, RGB + depth, and RGB + IR + depth. All experiments were evaluated on the test set, and the results are shown in Table 7.

From the perspective of precision and recall, the RGB-only model achieves 60.6% and 31.5%, respectively. After introducing the depth modality, P and R improve by 10.3 and 31.1 percentage points, respectively. After introducing the IR modality, the gains further increase to 14.9 and 35.6 percentage points, respectively. When IR and depth are simultaneously fused (RGB + IR + depth), P and R reach 91.4% and 88.3%, which are 30.8 and 56.8 percentage points higher than those of the single-modality baseline and represent the best overall performance. This indicates that, under the YOLOv8

architecture, depth information is particularly effective in improving recall and can stably alleviate missed detections. In terms of average precision, multimodal fusion also shows clear advantages. Compared with the RGB-only baseline, the depth-enhanced dual-modal model improves mAP@0.5 and mAP@0.5:0.95 by 26.7 and 17.2 percentage points, respectively; the IR-enhanced dual-modal model improves them by 30.5 and 22.8 percentage points, respectively; and the tri-modal model improves them by 48.8 and 36.3 percentage points, respectively. The RGB + IR + depth tri-modal fusion configuration again achieves the best performance, further demonstrating that the illumination robustness of IR data and the geometric constraints of depth data are complementary and, when fused with RGB images, can simultaneously improve classification correctness and localization precision under complex illumination and background interference. Since this conclusion holds consistently on both YOLOv5 and YOLOv8 baselines, the gains brought by the multimodal image data in the dataset are not dependent on a specific architecture and therefore exhibit strong generality.

6 Conclusion

This paper addresses two central challenges in multimodal object detection for driver behavior understanding: the insufficient robustness of single-modality vision schemes under complex illumination conditions and the severe lack of

Table 7 Detection results of different modality combinations based on YOLOv8

Backbone	Modality	AP (%)							P (%)	R (%)	mAP50 (%)	mAP50-95 (%)
		Head	Hand	Safety belt	Phone	Bottle	Cigarette	Steering wheel				
YOLOv8	RGB	90.4	35.0	28.0	12.8	19.2	16.3	90.8	60.6	31.5	41.8	28.9
YOLOv8	RGB + depth	99.5	87.4	59.7	39.6	41.6	52.5	99.5	70.9	62.6	68.5 (+26.7)	46.1 (+17.2)
YOLOv8	RGB + IR	99.5	90.2	73.4	61.5	47.2	35.1	99.5	75.5	67.1	72.3 (+30.5)	51.7 (+22.8)
YOLOv8	RGB + IR + depth	100.0	97.9	89.0	81.2	90.2	75.9	100.0	91.4	88.3	90.6 (+48.8)	65.2 (+36.3)

high-quality multimodal datasets with pixel-level alignment. To this end, we conduct a systematic study and construct CMMID, a large-scale intelligent-cockpit multimodal image dataset with diverse illumination conditions and precise pixel-level alignment, and establish CMMIR as a two-stage registration method that combines geometric mapping with mask-guided inpainting. The CMMID dataset contains synchronously acquired RGB, IR, and depth data and provides abundant instance annotations, thereby offering a solid data foundation for developing and validating high-performance multimodal perception algorithms. To verify the effectiveness of the dataset, we conduct a series of comparative experiments based on YOLOv5 and YOLOv8. The results demonstrate that the multimodal image data constructed in this work can effectively improve environment understanding in intelligent cockpit scenarios. Specifically, compared with the single RGB modality, the tri-modal configuration integrating IR and depth information achieves the best performance across all evaluation metrics. Using YOLOv5 as the baseline, the tri-modal fusion scheme improves mAP@0.5 by 50.6 percentage points and mAP@0.5:0.95 by 36.8 percentage points. The same trend also holds for YOLOv8, where mAP@0.5 and mAP@0.5:0.95 are improved by 48.8 and 36.3 percentage points, respectively. The analysis confirms that the illumination invariance of IR images effectively overcomes the detection difficulties caused by low-light and backlit environments, while the stable three-dimensional structural information provided by depth images substantially reduces missed detections and false detections caused by occlusion or background confusion. These modalities thus complement RGB information remarkably well. This work not only contributes a valuable benchmark dataset for multimodal perception in intelligent cockpit environments, but also experimentally verifies the effectiveness of multimodal fusion based on this dataset for improving the robustness and accuracy of driver monitoring systems. The proposed dataset also has broad potential applications in related tasks, such as driver behavior analysis, cockpit safety warning, and multimodal human-machine interaction.

Nevertheless, the current version of CMMID still has limitations in terms of sample diversity and scenario coverage. In future work, we will further expand the participant pool,

enrich the behavioral and scenario diversity of the dataset, introduce additional acquisition platforms and camera configurations, and continuously improve the evaluation protocols and baseline implementations, with the goal of establishing a more comprehensive and reproducible multimodal detection benchmark and supporting the exploration of more advanced fusion algorithms and models.

- [1] World Health Organization, *Global Status Report on Road Safety 2023: Country and Territory Profiles*. Geneva, Switzerland: World Health Organization, 2024.
- [2] D. McCarty and H. W. Kim, Risky behaviors and road safety: An exploration of age and gender influences on road accident rates, *PLoS One*, vol. 19, no. 1, p. e0296663, 2024.
- [3] J. Jang, C. Park, H. Kim *et al.*, Multispectral object detection enhanced by cross-modal information complementary and cosine similarity channel resampling modules, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, 2025, pp. 9437–9446.
- [4] O. Tamin, E. G. Mounq, J. A. Dargham *et al.*, On-shore plastic waste detection with YOLOv5 and RGB-near-infrared fusion: A state-of-the-art solution for accurate and efficient environmental monitoring, *Big Data Cogn. Comput.*, vol. 7, no. 2, p. 103, 2023.
- [5] T. Ophoff, K. Van Beeck, and T. Goedeme, Exploring RGB+depth fusion for real-time object detection, *Sensors*, vol. 19, no. 4, p. 866, 2019.
- [6] S. Gupta, R. Girshick, P. Arbelaez *et al.*, Learning rich features from RGB-D images for object detection and segmentation, in *Proc. Eur. Conf. Computer Vision*. Cham, Switzerland: Springer, 2014, pp. 345–360.
- [7] J. Kim and S. H. Baek, Pixel-aligned RGB-NIR stereo imaging and dataset for robot vision, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2025, pp. 11 482–11 492.
- [8] S. Song, S. P. Lichtenberg, and J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 567–576.
- [9] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, Multimodality image registration by maximization of mutual information, *IEEE Trans. Med. Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

- [10] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, Mutual-information-based registration of medical images: A survey, *IEEE Trans. Med. Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.
- [11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, UK: Cambridge Univ. Press, 2004.
- [12] J. Li, Q. Hu, and M. Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2019.
- [13] J. Li, P. Shi, Q. Hu *et al.*, RIFT2: Speeding-up RIFT with a new rotation-invariance technique, *arXiv preprint arXiv:2303.00319*, 2023.
- [14] M. P. Heinrich, M. Jenkinson, M. Bhushan *et al.*, MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration, *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, 2012.
- [15] F. Achermann, A. Kolobov, D. Dey *et al.*, MultiPoint: Cross-spectral registration of thermal and optical aerial imagery, in *Proc. Conf. Robot Learning*, 2021, pp. 1746–1760.
- [16] H. Le, F. Liu, S. Zhang *et al.*, Deep homography estimation for dynamic scenes, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 7652–7661.
- [17] Z. Teed and J. Deng, RAFT: Recurrent all-pairs field transforms for optical flow, in *Proc. Eur. Conf. Computer Vision*. Cham, Switzerland: Springer, 2020, pp. 402–419.
- [18] R. Suvorov, E. Logacheva, A. Mashikhin *et al.*, Resolution-robust large mask inpainting with Fourier convolutions, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, 2022, pp. 2149–2159.
- [19] W. Li, Z. Lin, K. Zhou *et al.*, MAT: Mask-aware transformer for large hole image inpainting, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 10 758–10 768.
- [20] G. Liu, F. A. Reda, K. J. Shih *et al.*, Image inpainting for irregular holes using partial convolutions, in *Proc. Eur. Conf. Computer Vision*, 2018, pp. 85–100.
- [21] A. Lugmayr, M. Danelljan, A. Romero *et al.*, RePaint: Inpainting using denoising diffusion probabilistic models, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.
- [22] A. Montoya, D. Holman, SF.data.science, T. Smith, and W. Kan, State farm distracted driver detection, [Online]. Available: <https://kaggle.com/competitions/state-farm-distracted-driver-detection>, 2016, accessed on: Feb. 9, 2026.
- [23] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, Real-time distracted driver posture classification, *arXiv preprint arXiv:1706.09498*, 2017.
- [24] M. Sabry, W. Morales-Alvarez, and C. Olaverri-Monreal, Automated vehicle driver monitoring dataset from real-world scenarios, *arXiv preprint*, 2024.
- [25] C. H. Weng, Y. H. Lai, and S. H. Lai, Driver drowsiness detection via a hierarchical temporal deep belief network, in *Proc. Asian Conf. Computer Vision*. Cham, Switzerland: Springer, 2016, pp. 117–133.
- [26] J. Wang, W. Li, F. Li *et al.*, 100-driver: A large-scale, diverse dataset for distracted driver classification, *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 7061–7072, 2023.
- [27] Driver activity dataset, [Online]. Available: <https://www.kaggle.com/datasets/guanhuallee/driver-activity-dataset>, accessed on: Feb. 10, 2026.
- [28] O. Kopuklu, T. Ledwon, Y. Rong *et al.*, Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework, in *Proc. 15th IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2020, pp. 77–84.
- [29] M. Martin, A. Roitberg, M. Haurilet *et al.*, Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles, in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2019, pp. 2801–2810.
- [30] J. D. Ortega, N. Kose, P. Canas *et al.*, DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis, in *Proc. Eur. Conf. Computer Vision*. Cham, Switzerland: Springer, 2020, pp. 387–405.
- [31] R. Zhao, B. Lu, X. Tang *et al.*, Multi-source hybrid-modality dataset and hierarchical fusion classification method for intelligent cockpits, *CAAI Trans. Intell. Syst.*, vol. 21, no. 0, pp. 1–12, 2026.
- [32] J. S. Katrolia, B. Mirbach, A. El-Sherif *et al.*, TICaM: A time-of-flight in-car cabin monitoring dataset, *arXiv preprint arXiv:2103.11719*, 2021.

Author biography



Xin Ning is a Professor and Ph.D. supervisor at the Institute of Semiconductors, Chinese Academy of Sciences. He is a Senior Member of IEEE, CCF, CAAI, and CSIG, and was listed among the world's top 2% scientists from 2022 to 2024. He is also a member of the Youth Innovation Promotion Association, Chinese Academy of Sciences. He has published more than 100 papers, has been granted more than 30 Chinese invention patents, and has authored one English monograph. He has served as Program Chair or Publication Chair for HDIS 2025, IEEE HPBD&IS 2021, Edutainment & Metaverse 2022, and ICCD 2023. He serves as an Area Editor for *Information Fusion* (Elsevier) and as an editorial board member of *IEEE TCSVT*, *Pattern Recognition*, *EAAI*, and *Journal of Intelligent Systems*. He has led or participated in projects including National Key R&D Programs, NSFC Young Scientist and General Programs, and Beijing Natural Science Foundation projects. He received the Second Prize of Science and Technology Progress from the Chinese Institute of Electronics, the First Prize of the First Youth Innovation Award of the Institute of Semiconductors, and was selected for the Young Researcher Program of the Institute of Semiconductors. E-mail: ningxin@semi.ac.cn.





tion, computer vision, and intelligent cockpit perception. E-mail: zhaokai@semi.ac.cn.

Kai Zhao received the B.Eng. degree from Beijing University of Chemical Technology, Beijing, China, in 2024, and is currently pursuing the master's degree at the Institute of Semiconductors, University of Chinese Academy of Sciences, Beijing, China. His research interests include multimodal fusion, computer vision, and intelligent cockpit perception. E-mail: zhaokai@semi.ac.cn.



Dan Li is a Master's Degree Candidate. Her primary research interests include computer vision, intelligent driving systems, AI-assisted healthcare, and EEG signal analysis and processing. She is a recipient of the National Scholarship. Additionally, she has been honored with the titles of Merit Student of Beijing and Outstanding Graduate of Beijing. To date, she has published 2 academic papers in renowned domestic and international journals and conferences, and has authored 1 book. Email: lidan251@mailsucas.ac.cn

Dan Li is a Master's Degree Candidate. Her primary research interests include computer vision, intelligent driving systems, AI-assisted healthcare, and EEG signal analysis and processing. She is a recipient of the National Scholarship. Additionally, she has been honored with the titles of Merit Student of Beijing and Outstanding Graduate of Beijing. To date, she has published 2 academic papers in renowned domestic and international journals and conferences, and has authored 1 book. Email: lidan251@mailsucas.ac.cn



Jin Ning received the B.S. degree in Electronic Engineering from Zhengzhou University, Zhengzhou, China, in 1995, the M.S. degree from the Lanzhou Institute of Physics, China Aerospace Science and Technology Corporation, Lanzhou, China, in 2000, and the Ph.D. degree from the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China, in July 2003. In 2003, he joined the Integrated Technology Center, Institute of Semiconductors, Chinese Academy of Sciences, where he has long been engaged in research on novel semiconductor microelectronic devices and MEMS (micro-electro-mechanical systems) devices. As a key member and principal investigator, he has completed and undertaken projects including the National Natural Science Foundation of China, the National 863 Program, and other provincial and ministerial projects. He has published more than 40 academic papers and delivered numerous conference presentations. E-mail: ningjin@semi.ac.cn

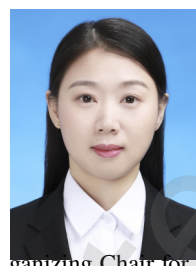
Jin Ning received the B.S. degree in Electronic Engineering from Zhengzhou University, Zhengzhou, China, in 1995, the M.S. degree from the Lanzhou Institute of Physics, China Aerospace Science and Technology Corporation, Lanzhou, China, in 2000, and the Ph.D. degree from the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China, in July 2003. In 2003, he joined the Integrated Technology Center, Institute of Semiconductors, Chinese Academy of Sciences, where he has long been engaged in research on novel semiconductor microelectronic devices and MEMS (micro-electro-mechanical systems) devices. As a key member and principal investigator, he has completed and undertaken projects including the National Natural Science Foundation of China, the National 863 Program, and other provincial and ministerial projects. He has published more than 40 academic papers and delivered numerous conference presentations. E-mail: ningjin@semi.ac.cn



Weijun Li received the Ph.D. degree from the Graduate University of the Chinese Academy of Sciences, Beijing, China, in 2004. He is currently a Professor and a doctoral supervisor. He is a Senior Member of the IEEE and the Chinese Association for Artificial Intelligence (CAAI), and serves as an Expert Committee Member of the China Integrated Circuit Intellectual Property Alliance. He is also a member of the Steering Committee of the International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), and served as the Organizing Committee Chair of HPBD&IS in 2019 and 2020. He has served as

Weijun Li received the Ph.D. degree from the Graduate University of the Chinese Academy of Sciences, Beijing, China, in 2004. He is currently a Professor and a doctoral supervisor. He is a Senior Member of the IEEE and the Chinese Association for Artificial Intelligence (CAAI), and serves as an Expert Committee Member of the China Integrated Circuit Intellectual Property Alliance. He is also a member of the Steering Committee of the International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), and served as the Organizing Committee Chair of HPBD&IS in 2019 and 2020. He has served as

a reviewer for several major journals, including *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Image Processing*, *IEEE Access*, *IEEE Signal Processing Letters*, *Neurocomputing*, *Science China*, and *CAAI Transactions on Intelligent Systems*. He has published more than 100 academic papers, authored one academic monograph, obtained more than 20 granted invention patents, and completed the technology transfer of more than 10 research outcomes. E-mail: wjli@semi.ac.cn



Baoli Lu, Ph.D., is an Assistant Professor at the Institute of Semiconductors, Chinese Academy of Sciences. She is a Senior Member of the China Computer Federation and a member of the Chinese Association for Artificial Intelligence. She has been a government-sponsored visiting scholar and served as Organizing Chair for IEEE HPBD&IS 2021 and IEEE HDIS 2022. Her long-term research focuses on fundamental theories and key technological applications in computer vision, intelligent systems, and AI-assisted diagnosis and treatment. As a sub-project leader and key project member, she has participated in National Key R&D Programs, NSFC projects, and Beijing Natural Science Foundation projects. She has published more than 20 papers in leading domestic and international journals and conferences, applied for more than 20 invention patents, and has been granted 10 patents. She has also won multiple algorithm competition awards. E-mail: lubaoli@semi.ac.cn.

Baoli Lu, Ph.D., is an Assistant Professor at the Institute of Semiconductors, Chinese Academy of Sciences. She is a Senior Member of the China Computer Federation and a member of the Chinese Association for Artificial Intelligence. She has been a government-sponsored visiting scholar and served as Organizing Chair for IEEE HPBD&IS 2021 and IEEE HDIS 2022. Her long-term research focuses on fundamental theories and key technological applications in computer vision, intelligent systems, and AI-assisted diagnosis and treatment. As a sub-project leader and key project member, she has participated in National Key R&D Programs, NSFC projects, and Beijing Natural Science Foundation projects. She has published more than 20 papers in leading domestic and international journals and conferences, applied for more than 20 invention patents, and has been granted 10 patents. She has also won multiple algorithm competition awards. E-mail: lubaoli@semi.ac.cn.