Non-intrusive Automatic 3D Gaze Ground-truth System

Feng Hu, PhD

Autonomous Vehicle Software, NVIDIA (fengh@nvidia.com)

Abstract: Driver distraction has surfaced as a significant safety issue worldwide, and the capacity to track a driver's attention via monitoring its gaze direction is one of the most critical features in the modern Driver Monitoring System (DMS). Deep learning based gaze estimation has grown in popularity due to its robustness across operating conditions. Though appropriate network structure design and parameters tuning are important, accurate ground-truth estimation for millions of gaze training images to build the model also plays a critical role in achieving high-quality gaze estimation results. This paper proposes a non-intrusive automatic 3D ground-truth data collection system for large-scale on-bench and in-car data collection, using gamified camera calibration, occlusion invariant mirror-based camera localization, and noise-robust 3D reconstruction algorithms. Experimental results are provided to demonstrate the system's accuracy and robustness even in challenging conditions.

CCS CONCEPTS • Computing methodologies • Artificial intelligence • Computer vision

Keywords: Driver Monitoring System (DMS), Gaze estimation, Camera calibration, Ground-truth system

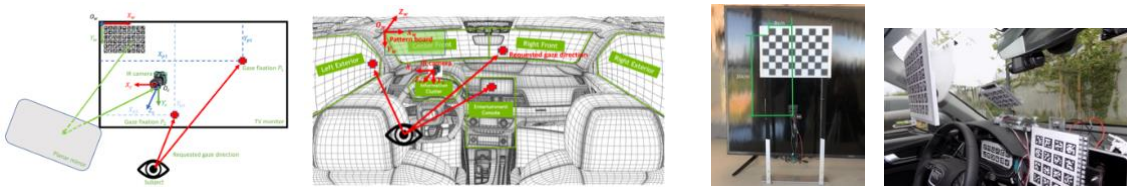
1 INTRODUCTION

Driver Monitoring System (DMS) uses gaze direction information to determine what the driver of a vehicle is viewing at a given point in time. For example, is the driver looking at the front windshield or elsewhere in the car, such as the rear-view mirror or instrument panel? This information can be helpful in further determining whether a driver is distracted or unable to see an action happening around a vehicle, such as a pedestrian walking in front of the car. If the DMS determines that a driver is distracted and may not see the pedestrians, it can stop the vehicle autonomously to avoid striking them.

Previous researchers use intrusive mobile eye-tracking glasses [1][2] to obtain the ground-truth for gaze estimation, which are not preferred for an end-to-end Deep Neural Network (DNN) solution since the extra glasses partially occlude the face and are detrimental to the gaze accuracy. Furthermore, even though semantic image inpainting to the area covered by the glasses can be applied to remove the obtrusiveness, the effect is limited, and a non-intrusive way of establishing ground-truth without occlusion is preferable.

We collect DMS gaze data in both indoor and outdoor environments, such as in front of an on-bench TV display or inside a real car, as shown in Fig. 1. The on-bench setup is easier to build, cheaper to scale up and generally allows faster data collection since each subject consumers less time. However, it is not sufficient if we need data about a subject's looking beyond a TV display especially mimicking a real driver's gaze movement, or if we need gaze testing data while a vehicle is moving. In these cases, we require an in-car setup.

The camera types for gaze data collection and DNN model deployment (i.e., inference) may differ. For example, we can collect data on the bench using an RGB camera and deploy the model in a car using an InfraRed (IR) camera. In this paper, we conceptualize a virtual camera that only contains extrinsic parameters (i.e., rotation and translation) without intrinsic constraints. We use it as a bridge to connect data collection setups and possible deployment setups. In addition, we define the position where a subject is asked to stare at as a gaze *fixation point*. Therefore, the task of 3D ground-truth for a given gaze image is defined as calculating the 3D coordinates of the fixation point in the virtual camera space.



(a) On-bench and in-car setup diagram (b) real on-bench and in-car setup images

Figure 1: DMS gaze data collection diagram and real images

We need the DMS gaze estimation to work day and night in challenging environments such as under strong sunshine. Regular passive RGB cameras will not work in these environments due to lacking illumination or over-exposure. Our work mainly focuses on Infrared (IR) camera-based systems using active LED-based illumination.

We summarize our contributions as follows: (1) a non-intrusively automatic large-scale DMS gaze data collection system for both on-bench and in-car setups without requiring a subject to wear any special equipment; (2) a new 3D ground-truth concept for each gaze image defined in virtual camera space and a coordinate propagation mechanism that enables calculating the ground-truth for very dense gaze fixation points; and (3) a gamified camera calibration, an occlusion invariant mirror-based camera localization and a noise-robust 3D reconstruction method for accurate and robust 3D ground-truth calculation.

The rest of the paper is organized as follows. First, related work is discussed in Section 2. Next, the methodology is described in Section 3, and experimental results are shown in Section 4. We finally conclude our paper and discuss our limitations as well as dataset contributions in Section 5.

2 RELATED WORK

Gaze estimation has many use cases in various platforms, such as desktops, monitors, handheld devices, head-mounted devices, or automotive [3]. In this paper, we focus on automotive scenarios, where we use gaze to estimate where drivers or passengers are looking to evaluate their vigilance or distraction levels.

In recent years, using Deep Learning (DL), such as Convolutional Neural Network (CNN), to build direct end-to-end gaze estimation solution has attracted more and more attention [4] due to its robustness across operating conditions. However, many gaze images with accurate gaze ground-truth labels are necessary for training robust gaze models. To obtain the gaze direction vector ground-truth, we need two coordinates, the eye's location and the location of the point where the eye is looking. Previous researchers use pseudo-ground-truth labels [5] or complex hardware setups (i.e., 18 cameras, projector, and light boxes) [1] to create ground-truth labels. However, these works cannot satisfy the demands of a DMS/OMS environment, which

requires real ground-truth and a broader region of interest (ROI). This paper will focus on real 3D ground-truth generation for providing DNN with gaze images and ground-truth labels using minimal hardware requirements.

Traditional camera calibration is a well-studied problem where the camera directly perceives a target pattern board, and we take images to estimate the camera's intrinsic and extrinsic parameters [6-8]. Camera extrinsic calibration (or camera localization) regarding a pattern board not inside the camera's Field of View (FoV) is a challenging problem. Still, it frequently occurs in gaze estimation applications on monitors, handheld devices (such as smartphones), or automotive platforms [9]. Researchers employ mirrors and chessboards to estimate the relative posture and position of the camera (i.e., extrinsic parameters) against a 3D reference object not directly visible to the camera [10]. However, we use active illumination and Infrared (IR) cameras in automotive cases. Mirrors can reflect away LED illumination causing higher signal-noise ratio and corner detection difficulty, making the overall calibration process very challenging. In this paper, we propose a robust occlusion-invariant AprilTag-based camera localization approach to address these problems for IR cameras.

To support 3D fixation point data collection, we need an additional 3D reconstruction process. Previous research uses AprilTags for offline reconstruction [11]. However, many tags on one AprilTag board may be missing due to noise and occlusion. In our work, we design a RANSAC [12] based tag recovering approach to improve the accuracy and robustness of the 3D reconstruction process.

3 METHODOLOGY

Training an accurate and robust DNN model that works under various conditions requires a significant amount of training data and respective ground-truth labels. This section presents how we define and calculate the gaze ground-truth label for a given gaze image at a large scale.

We define a gaze image's 3D ground-truth as the coordinates of the fixation point in the camera coordinate system. By requesting a subject to look at various fixation points on-bench or in-car, we collect many gaze frames or short video clips for this subject. The exact process can be repeated to many subjects until enough gaze training data are collected, though the definition of sufficiency varies depending on multiple factors such as network structure or application specifications. Since usually large quantities, e.g., in millions, of training images are needed, we use multiple cameras and record images simultaneously to speed up the data collection process.

There are multiple advantages to an in-car setup. First, it can collect data for regions hard for the on-bench setup to provide, e.g., the right front passenger window area. Second, it can give a testing image dataset for the in-car evaluation of a trained gaze DNN model.

3.1 On-bench Setup and Camera Localization

We obtain gaze images of subjects and their respective 3D ground-truth with our indoor on-bench setup. In the on-bench case, only camera localization is required since all fixation points are on the same plane, and no 3D reconstruction is needed.

3.1.1 On-bench setup

The workflow of collecting one image and ground-truth pair on-bench is as follows. First, one or multiple cameras are installed on a TV display surface, facing toward a subject sitting on a chair. Then, the subject is

requested to stare at a marker (such as a "+") shown on display for a while. Next, a mouse is provided for left-or-right-clicking to notify the system that the subject has started or finished looking at a fixation point. After that, the system shifts the fixation point to a new location, and the same procedure is repeated until we collect enough data for this subject.

Since the ground-truth of a fixation point is defined as the point's coordinates in the camera coordinate system, we need to calculate the geometric relationship between the world coordinate system and the camera coordinate system, i.e., the relationship between the TV monitor plane and the IR camera. We denote this procedure as camera localization since we determine the camera's location in the world coordinate system. It is a special case of extrinsic camera calibration where the world coordinate system origin is not inside the camera's Field of View (FoV).

The approach for camera localization is different from the traditional camera extrinsic calibration. In regular extrinsic calibration, the camera can see the entire pattern board where the world coordinate system is defined, and Zhang's method [6] can be directly applied. In contrast, in camera localization, we need a particular way to enable the camera to see the pattern board, typically using a mirror reflection.

There are multiple challenges for IR camera localization compared with regular RGB camera localization. First, the mirror will redirect away some of the active illumination generated by surrounding LEDs, making the chessboard too dark and noisy to detect fully. Second, a part of the chessboard may be occluded by in-vehicle obstacles, such as the steering wheel, which can fail the camera localization process.

3.1.2 AprilTag-based camera localization

We propose an occlusion-invariant Apriltag-based camera localization approach to overcome the IR camera problems. Even if some of the tags are occluded or too noisy to be detected, the camera localization process will still work. However, in challenging situations, such as under poor illumination or having strong sunshine, the algorithm must complete the localization process only using the detected part of the Apriltags.

The left image of Fig. 2 illustrates the geometry diagram of using a single mirror π_i to reflect multiple AprilTag points. Denote P_{real}^1 and P_{real}^2 as two real fixation points located on a 2D plane. They are reflected by an arbitrary mirror j and generate two virtual points P_j^1 and P_j^2 , which are then captured by an IR camera and imaged as p_j^1 and P_j^2. Note that the superscript i of P_j^i is the variable to index 3D points, and the subscript j is to index different mirror poses. When the mirror moves to a different position, we treat the new position as a new mirror pose. If j equals $real$, it means the point is on the real TV display rather than a mirrored virtual point. Uppercase, e.g., P is used to indicate a 3D point, and its correspondent lowercase, e.g., p, is used to denote the correspondent 2D image point. The middle image of Fig. 2 illustrates the geometry diagram of multiple mirrors plane reflecting the same point and making their virtual points to be inside the camera FoV and imaged.

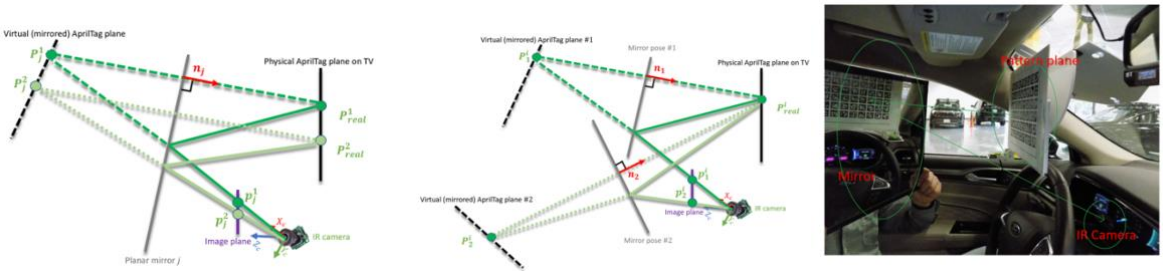


Figure 2: Reflecting multiple points using a single mirror (left), reflecting a single point using different mirror poses (middle), and sample camera and pattern relationship in-car (right)

3.1.3 Building the camera localization equation system

Representing the camera localization result with a rotation matrix R and a translation vector T. These two parameters convert any fixation point's world coordinates into camera coordinates using Equation 1, where P_w is a fixation point's 3D coordinates in the world coordinate system, and P_c is its correspondent coordinates in the camera coordinate system, i.e., the ground-truth.

$$Pc = R \cdot Pw + T \quad (1)$$

We observe that the AprilTag points' coordinates in the camera coordinate system can be calculated in two ways. First, according to Equation 1, they can be represented using its known world coordinate system and R as well as T, as shown on the left side of Equation 2. Second, they can be represented by leveraging their mirrored points using Household transformation, as shown on the right side of Equation 2.

$$R \cdot P^i_{real} + T = -2\left(n_j^T \cdot P_j^i + d_j\right) \cdot n_j + P_j^i \quad (2)$$

We will explain a little more on the right side of Equation 2. First, using the Perspective-n-Points (PnP) algorithm [13], we can calculate any virtual points P_j^i 's coordinates in the camera coordinate system, since the distances in-between the points are known.

Assume no two mirror poses are parallel with each other in practice, which is feasible since we guided the user to move the mirror into our targeted poses, a sample image of which is shown in Fig. 3. Therefore, any pair of planes defined by a pair of mirror poses will always intersect with each other and have exact one intersection line. In addition, if one mirror pose exists in at least two mirror pose pairs, and the two respective intersection lines within the mirror plane can be calculated, we can further apply vector cross product on these two intersection lines to calculate the mirror plane's normal vector.

Now to calculate a mirror's normal vector, we only need to figure out how to calculate intersection lines within the mirror plane. Denote the intersection line defined by mirror pose #1 and #2 in Fig. 4 as L_{12}. By mirror reflection law, we know vector $P_1^i P_{real}^i$ is perpendicular to every line within mirror plane #1 including L_{12}, and $P_2^i P_{real}^i$ is perpendicular to every line within mirror plane #2, including L_{12}. Therefore, L_{12} is perpendicular to each line on the plane defined by vector $P_1^i P_{real}^i$ and $P_2^i P_{real}^i$, including line $P_1^i P_2^i$. Since both P_1^i and P_2^i are known using the PnP algorithm, we can calculate $P_1^i P_2^i$'s normal vector L_{12}. Repeating this procedure to at least one more intersection line, e.g., intersection line L_{13} with a third mirror pose #3, we can figure out mirror pose #1's normal vector as $L_{12} \times L_{13}$, where × is the vector cross-product operator.

Finally, we can establish a system of equations to solve R, T, and distance d_j together using Equation 2, and use nonlinear optimization, such as Levenberg–Marquardt algorithm, to optimize R, T, n_j and d_j together.

From the estimated rotation matrix R and translation vector T, which transforms a point in the world coordinate system P_w into a camera coordinate system P_c, we can estimate the camera origin's distance to the world coordinate system with Equation 3, which can be used as one of the criteria to evaluate localization quality.

$$d = -R^{-1} \cdot T \quad (3)$$

3.1.4 Comparison with previous work

We have multiple innovations compared with previous work, e.g., reference [10]. First, [10] only uses the chessboard pattern and requires all chessboard corners to be fully detected as a prerequisite to make camera localization work. However, almost every DMS uses an active illumination-based infrared camera to cover day and night usage, and it's incredibly challenging, if possible, to guarantee all chessboard corners can be detected in practice due to insufficient or unbalanced illumination after mirror reflection, especially for multi-cameras setups. Moreover, occlusion from in-vehicle objects such as the steering wheel or slightly larger FOV lens' distortion will make it even harder for the prerequisite to be valid. In our work, because our approach is occlusion invariant, it significantly eases the prerequisite by only requiring part of the corners detected, e.g., 70%, via using distributed AprilTag, which substantially increases the usability making the approach to be useful from theory to a real industrial environment. Finally, we optimize each camera's extrinsic parameters and all mirror poses online without requiring knowing beforehand the mirror poses number and detected AprilTag corner number, even if whether a tag can be detected by a mirror is unpredictable, which does not appear in previous work.

3.2 In-car Setup and 3D Reconstruction

Subjects sit in the driver's seat and are requested to look at pre-defined fixation points inside or outside of the car one by one to collect gaze training and testing images in a vehicle. The IR camera is installed in front of the subject, e.g., under the dashboard, and captures the subject's high-resolution face images. A sample IR camera installation is shown in the right image of Fig. 3.

There are two challenges for the in-car setup. First, fixation points can no longer be shown electronically, e.g., as a marker on a TV display. Therefore, a new mechanism is needed to hint the subject where to stare. To solve this problem, we designed a Raspberry Pi controlled LED panel mounting at various fixation areas and controlling the LED's on or off status using their indexes on the board. The subjects are then requested to look at the LED when it is turned on to provide gaze data. The white square board behind the right bottom AprilTag board shown in the right image of Fig. 1(b) is a sample LED board.

Second, for on-bench setup, all the gaze fixation points are on the same plane, e.g., a TV display plane; however, for in-car setup, the fixation points are not necessarily coplanar, e.g., the three points marked with a red crossing in the right image of Fig. 1(a). To solve this problem, we first designed a pattern mount installed inside the cabin, as shown in the right image in Fig. 2, serving the TV plane's role in the on-bench setup. Then we create a robust 3D reconstruction process to calculate each LED's coordinates in the world coordinate system and integrate the camera localization results and reconstruction results to complete ground-truth calculation.

3.2.1 In-car Camera Localization

To localize the camera, a user holds a mirror and reflects the pattern board back into the IR camera's field of view. This step is repeated until enough images, e.g.,20, are collected. Camera localization results can be instantly calculated after images are captured, e.g., within one or two minutes, and stored automatically for usage.

There are multiple factors to guarantee a successful camera localization. First, appropriate LED numbers and installation locations are essential. IR cameras have frequency filters, and almost all the illumination is from the active LEDs. Second, the mirror only reflects part of the rays back into camera FoV and may cause the images to be dark and blurry. We recommend using two LEDs and mounting them on the left and right sides of the camera. In addition, for some vehicle models, the steering wheel may partially block the camera's FoV or the rays reflected by the mirror, so adjusting the steering wheel before the camera localization to maximize the total number of visible AprilTags can be helpful.

3.2.2 In-car 3D Reconstruction

Neural network training generally prefers the input data to be well distributed to avoid biased sampling. In the DMS gaze case, we expect a rich amount of gaze fixation points distributed, preferably not just on one plane. In addition, we expect the spatial resolution of the fixation points to be dense enough for training high-accuracy gaze models. To satisfy these requirements, we design a set of LED panels and overlay thin AprilTag boards on top of them to provide high spatial resolution gaze fixations and use in-car 3D reconstruction to calculate their ground-truth.

We set the distance between two adjacent LEDs to 1cm. Given that the subject usually sits more than 50cm away from the fixation points, this distance can provide sufficient spatial resolution for most gaze specifications. A thin AprilTag board is pasted and aligned on top of the LED boards, as shown in the right image of Fig. 2, so the geometric relationship between the LED board and the AprilTag board can be measured. In other words, if an AprilTag's coordinates in the world coordinate system are known, the LEDs' coordinates can be estimated.

Finding the ground-truth of each LED is then converted into the problem of localizing each AprilTag board in the world coordinate system. Previous work uses Structure-from-Motion (SfM) [14] to estimate each AprilTag board's coordinates. However, due to illumination noise, some of the tags on AprilTag boards may be missed or drifted in detection. In our work, we first use RANdom SAmpling and Consensus (RANSAC) [12] to recover the missed tags and remove the outliers. Doing so allows us to utilize most of the detected AprilTags for a more robust and accurate reconstruction rather than relying on one single tag.

4 EXPERIMENTS

The hardware requirements for conducting experiments on both on-bench and in-car setups are as follows. For on-bench setup, a TV monitor connecting to a computer, a mirror, a printed pattern board, one or multiple IR cameras, and their illumination LEDs are needed. For in-car setup, we need a calibration hardware mount, one or multiple IR cameras and their illumination LEDs, a mirror, and a set of Raspberry Pi controlled LED boards.

Note that the mirrored AprilTag will be different from regular AprilTag, and the existing detection algorithm will not be able to recognize such tags in an image. To solve this problem, rather than directly printing the tags

on the board, we print a pre-mirrored set of AprilTag to cancel the reflection effect. Then, when this set of tags is reflected again by the mirror, the camera will see normal AprilTags, and we can detect the tags without modifying the AprilTag detection algorithms.

4.1 Camera Intrinsic Calibration

For both setups, intrinsic camera parameters are needed for later localization tasks. However, regular chessboard-based camera intrinsic calibration using Zhang's approach [6] may have problems. For example, degenerated cases such as feeding the calibration algorithm with the same pose image many times by users without domain knowledge can happen. To resolve this problem, we gamified the process by designing a set of pre-defined poses and guiding the users to adjust the pattern board to fit the recommended pose. The left image of Fig. 3 shows a sample suggested pattern pose where the arrows can guide the user to adjust the current board's posture.

We establish criteria to ensure the best possible calibration results, including reprojection error threshold, input image quality check, sufficient chessboard image number, variance check, and result visualization.

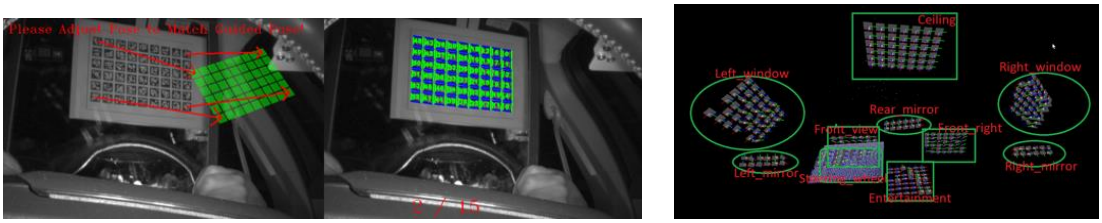


Figure 3: Camera localization example (left) and vehicle 3D reconstruction result sample (right)

In our experiments, we set input images to 20 and design a set of various poses to guide the users to collect the images accordingly. In addition, the reprojection threshold is set to 2 pixels for a successful calibration, and we automatically reject a trial if the average reprojection error is larger than the threshold.

4.2 Camera Localization

Three criteria are established to evaluate the camera localization quality. First, reprojection error is helpful in fast reject failures similar to camera intrinsic calibration. In our experiments, localization results with an average reprojection error more significant than 2 pixels are rejected instantly. Second, using Equation 3, we can calculate the estimated camera origin's distance to the world coordinate origin, which can also be measured using a tape. If the distance difference between the estimated and measured results is more than a threshold, e.g.,2cm, the localization trial is rejected. Third, multiple localization trials can be conducted for a given setup, and we can set a variance threshold to refuse flawed localization trials. We accept the camera localization result only when all three criteria are satisfied.

The left image of Fig. 3 shows an example of image capture in camera localization using a mirror and an AprilTag board. Even though all AprilTags are detected in the image, this is not required, and in many challenging cases, such as extreme mirror poses or occlusion, it is impossible to catch all tags. In our work, We set a configurable occlusion ratio, e.g., less than 70%, to control the acceptance of occlusion, which significantly increases the solution's usability.

Localization result variance is another important criterion for providing quality confidence evaluation. The variance of a set of localization experiments is defined in Equation 4

$$V_{loc} = \sqrt{\sum_{i=1}^{N} (f_{norm}(C_i - C'))^2} \quad (4)$$

where C_i is the estimated camera center coordinates in the world coordinate system using Equation 3, C' is the geometric camera center of all the trials, f_{norm} is the Euclidean distance operator between two 3D vectors, and N is the number of trials.

Table 1 shows an example of detailed camera localization results with 5 different trials, where each row provides a pair of rotation and translation results for one test. In this experiment, the variance is 0.36cm smaller than the threshold; therefore, the ground-truth system accepts the results.

4.3 3D Reconstruction

This subsection will present the quality evaluation criteria for reconstruction and some sample reconstruction results.

To obtain a gaze image's ground-truth, we need the camera localization procedure to transform the fixation point's coordinate from the world coordinate system to the camera coordinate system. In addition, we also need the AprilTag-based 3D reconstruction to estimate the point's coordinates in the world coordinate system, given its LED index.

Multiple criteria are set to evaluate the reconstruction quality and reject the trials that perform below the thresholds. First, the reprojection error of the Structure from Motion is checked to ensure the reconstruction is completed and satisfies the threshold, e.g., a few pixels. Second, we select a few AprilTag points as control points and measure their inter-distances. These distances are then compared with the estimated distance using Equation 1. We accept the reconstruction results when the difference is within a small range, e.g.,1-2 cm. Third, if the fixation points are on some key locations whose 3D coordinates are known from vehicle CAD models, we can compare the reconstructed results with the manufacturer CAD results and reject or accept based on customized threshold settings.

The right image of Fig. 3 shows an example of the 3D reconstruction results. Since the reconstruction process generates a rotation matrix and a translation vector for each AprilTag board, we can use them to visualize the boards' poses, as shown in the figure. In addition, these visualization results can be further visually checked to ensure they are consistent with the actual locations where the boards are placed.

In this result, each green marker shows one board reconstructed pose in 3D space. Some of the tags, such as the ones in the left window board, are missing due to poor image quality, e.g., blurring; however, the board is still reconstructed using the detected tags. In addition, a visual check is conducted to ensure that the estimated tag board topology is the same as the actual board topology inside the physical vehicle.

Table1 Camera localization multiple trials results (units: rad and meter)

$Euler_x$	$Euler_y$	$Euler_z$	t_1	t_2	t_3
2.865	0.046	-3.105	10.759	21.079	-12.281
2.862	0.042	-3.105	10.690	20.999	-12.209
2.780	0.029	-3.096	10.714	21.068	-11.779
2.862	0.052	-3.104	10.735	20.987	-12.086
2.998	0.068	-3.107	10.745	21.054	-12.072

5 CONCLUSION

We propose a non-intrusive automatic Driver Monitoring System (DMS) gaze 3D ground-truth data collection solution for both on-bench and in-car physical setups. We establish the 3D ground-truth concept inside a virtual camera coordinate system and design a coordinate propagation mechanism to make 2D TV plane points as well as 3D points (inside or outside of a vehicle) ground-truth possible. Finally, we develop a gamified camera calibration method, an occlusion-invariant mirror-based camera localization approach, and a noise-robust 3D reconstruction algorithm to enable accurate and highly usable 3D ground-truth calculation.

Limitation and dataset contribution. Gaze datasets include sensitive privacy data, e.g., very high-resolution full-face images from multiple angles; therefore, it is hard to make publishing such datasets legally feasible. However, we make collecting such a dataset nonintrusively clear, so any person or organization can repeat this method to collect their data for academic or industrial usage. In addition, due to space limits, this paper only focuses on the data collection system and does not contribute to the gaze Deep Neural Network (DNN) design and training part. However, we have published our pretrained gaze DNN models and related details using our collected dataset. They are publicly available in the “Gaze Estimation” section at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tao/models/gazenet.

REFERENCES

[1] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In European Conference on Computer Vision, pages 365–381. 2020.

[2] Diederick C Niehorster, Roy S Hessels, and Jeroen S Benjamins. Glassesviewer: Open-source software for viewing and analyzing data from the tobii pro glasses 2 eye tracker. Behavior Research Methods, 52(3):1244–1253, 2020.

[3] Feng Hu, Niranjan Avadhanam, Yuzhuo Ren, Sujay Yadawadkar, Sakthivel Sivaraman, Hairong Jiang, Siyue Wu. Gaze detection using one or more neural networks. United States Patent US 11,144,754. United States Patent and Trademark Office. Oct. 2021.

[4] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. IEEE transactions on pattern analysis and machine intelligence, 2019.

[5] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. IEEE Conference on Computer Vision and Pattern Recognition, pages 9980–9989, 2021.

[6] Zhengyou Zhang. A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence, 22(11):1330–1334, 2000.

[7] Yuzhuo Ren, Feng Hu. Camera Calibration with Pose Guidance. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2180-2184, June 2021.

[8] Feng Hu, Yuzhuo Ren, Niranjan Avadhanam, Ankit Pashiney. System and Method for Optimal Camera Calibration. United States Patent US 2021/010575. United States Patent and Trademark Office. April. 2021.

[9] Christian Nitschke, Atsushi Nakazawa, and Haruo Takemura. Display-camera calibration from eye reflections. In 2009 IEEE 12th International Conference on Computer Vision, pages 1226–1233. IEEE, 2009.

[10] Kosuke Takahashi, Shohei Nobuhara, and Takashi Matsuyama. A new mirror-based extrinsic camera calibration using an orthogonality constraint. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1051–1058. IEEE, 2012.

[11] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In 2011 IEEE International Conference on Robotics and Automation, pages 3400–3407. IEEE, 2011.

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.

[13] Francesc Moreno-Noguer, Vincent Lepetit, and Pascal Fua. Accurate non-iterative o(n) solution to the pnp problem. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE, 2007.

[14] Frank Neuhaus, Stephan Manthe, and Dietrich Paulus. Practical calibration of actuated multi-dof camera systems. In 8th International Conference of Pattern Recognition Systems (ICPRS 2017), pages 1–6, 2017.