

## RESEARCH ARTICLE

# Head Pose-Aware Regression for Pupil Localization From A-Pillar Cameras

**DONGHWA KANG**  **AND DONGWOO KANG** , (Member, IEEE)

School of Electronic and Electrical Engineering, Hongik University, Seoul 04066, South Korea

Corresponding author: Dongwoo Kang (dkang@hongik.ac.kr)


This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant 2022R1F1A1074056.

**ABSTRACT** In vehicular applications, remote eye pupil tracking is essential, particularly for advanced augmented reality (AR) 3D head-up displays (HUDs), and driver monitoring systems (DMS). However, achieving accurate pupil center localization under varied head poses presents significant challenges, especially when cameras are positioned on a vehicle's A-pillar. This placement introduces substantial head pose variations, complicating traditional tracking methods. In response, this study presents a remote pupil localization method designed to address the unique challenges posed by a camera situated on a vehicle's A-pillar, a spot causing significant head pose variations. The proposed technique relies on a head pose-aware pupil localization strategy utilizing A-pillar cameras. Our pupil localization algorithm adopts a Transformer regression approach, into which we integrate head pose estimation data, enhancing its capability across diverse head poses. To further enhance our approach, we used an optimized nine-point eye-nose landmark set, to minimize the pupil center localization loss. To demonstrate the robustness of our method, we conducted evaluations using both the public WIDER Facial Landmarks in-the-wild (WFLW) dataset and a custom in-house dataset focused on A-pillar camera captures. Results indicate a Normalized Mean Error (NME) of 2.79% and a failure rate (FR) of 1.28% on the WFLW dataset. On our in-house dataset, the method achieved an NME of 2.96% and a FR of 0.72%. These impressive results demonstrate the robustness and efficacy of our method, suggesting its potential for implementation in commercial eye tracking systems using A-pillar mounted cameras, especially for AR 3D HUD and DMS applications.

**INDEX TERMS** Pupil center localization, remote eye tracking, head pose-aware pupil regression, eye-nose points regression, head pose estimation, A-pillar camera, augmented reality (AR) 3D head-up displays (HUDs), driver monitoring system (DMS).

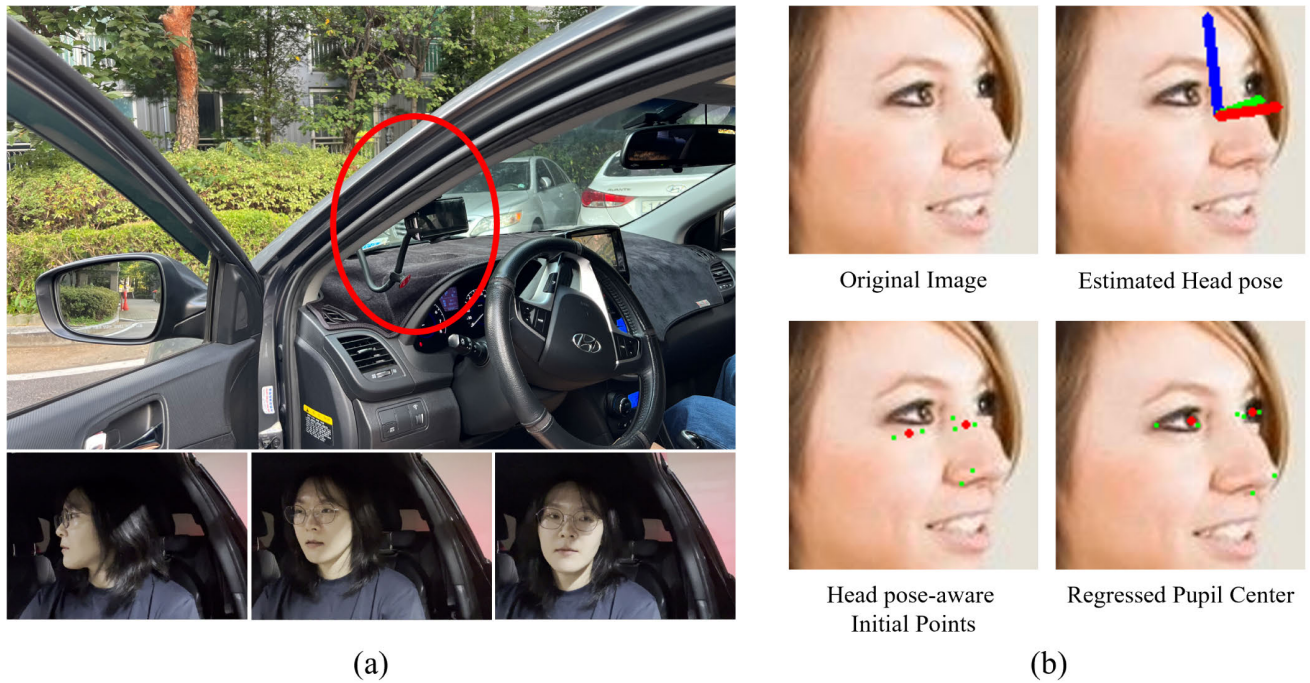
## I. INTRODUCTION

Pupil center localization plays an important role in many applications designed for drivers inside vehicles. One of the innovative vehicular user convenience systems is the combination of augmented reality (AR) and autostereoscopic three-dimensional (3D) display techniques in next-generation head-up displays (HUDs) [1]. These advanced AR 3D HUDs can show augmented reality 3D objects in line with the road, without the limitations of viewing zone boundaries [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Shajulin Benedict .

To provide users with a 3D visual experience free from 3D crosstalk and without the necessity of special 3D glasses, it's essential to use technology that accurately tracks the user's pupil center [3]. Additionally, in the area of driver monitoring systems (DMS), it is very important to assess a driver's alertness, concentration, and overall driving status [4]. In these DMS systems, eye-gaze information is essential, and pupil center localization is a key component of the process [5].

Eye tracking has traditionally been used in wearable AR and VR devices to mitigate 3D fatigue [6], [7] and enable effective human-computer interaction (HCI) [8]. Most



**FIGURE 1.** (a) Top: Illustration of the camera location at the A-pillar in vehicles. Bottom: Varied facial poses captured from our in-house dataset, showing the impact of this camera placement. (b) A step-by-step demonstration of the head pose-aware Transformer-based pupil center regression method on a public dataset, WFLW. Red dots indicate the pupil centers and green dots indicate other eye-nose points.

research and development in this area focus on using near-infrared (NIR) cameras, especially in AR/VR wearable glasses that are worn closely [9]. In contrast to this, our work on pupil localization aims to identify the pupil center from a distance. Remote pupil tracking has been useful in various applications, including autostereoscopic 3D displays [10], AR 3D HUDs [3], DMS [5], and HCI [11]. It's designed to quickly and accurately find the user's eye position from approximately 1 meter away, ensuring stable performance even if the eyes are occluded. However, these technologies usually expect the camera to be directly in front of the user's face and have difficulties with varying poses.

In commercial vehicles, directly placing the camera in front of the driver is highly challenging. This is because it would obstruct the driver's forward view, leading to potential hazards [12]. The automotive industry, aware of this, is now moving towards installing cameras on the A-pillar [13]. This setup ensures that the driver's view remains clear while still allowing for effective eye tracking. However, having the camera on the A-pillar presents challenges due to significant pose variations, making it difficult to predict pupil tracking using traditional models. This paper aims to enhance remote pupil localization performance, especially when the camera is set on the A-pillar and the driver shows pronounced head movements. Our method estimates a driver's head pose and incorporates it into a head pose-aware regression for pupil localization. Figure 1a shows a camera

on the A-pillar capturing the driver's faces, while Figure 1b provides an overview of our proposed head pose-aware pupil localization technique. The key contributions of our work include:

- **Head Pose-Aware Pupil Localization:** We introduce a novel approach that combines head pose estimation with pupil localization, specially designed for A-pillar mounted cameras in vehicles. This method effectively addresses the challenge of diverse head movements, ensuring accurate detection of the pupil center in different driving situations.

- **Transformer-Based Algorithm Integration:** Our work advances the state-of-the-art by incorporating head pose information into a Transformer-based regression algorithm. This integration enhances the algorithm's capability to accurately localize the pupil in varied conditions.

- **Optimized Nine Eye-Nose Points:** We propose a refined set of nine key eye-nose points, specifically optimized for remote pupil localization. This focus on essential keypoints, rather than the entire facial structure, reduces alignment errors and improves localization accuracy.

- **Comprehensive Evaluation with In-House Dataset:** In addition to the public face dataset, we tested our method with our own unique in-house dataset. This dataset captures real driving scenarios from an A-pillar camera. This two-fold dataset approach highlights our method's effectiveness across varied situations.

## II. RELATED WORKS

### A. FACIAL KEYPOINTS ALIGNMENT FOR PUPIL LOCALIZATION

The process of determining the pupil's center heavily depends on the alignment of facial landmarks. Initially, this process involves detecting the face using a face detector [14], [15], followed by aligning facial landmarks within the detected face area using techniques like face keypoints alignment [16]. The final step is localizing the pupil's position. Early methods relied on basic machine learning algorithms like Active Appearance Models (AAMs) [17] and other handcrafted feature-based methods, such as the Supervised Descent Method (SDM) [18], which used features like scale-invariant feature transform (SIFT) [19].

Deep neural networks (DNN) have significantly enhanced these methods, which are generally divided into two types: coordinate regression models and heatmap regression models. Coordinate regression models [15], [20], [21], [22] directly transform input features into vector coordinates of keypoints. For example, the TCDCN [15] performs multi-task learning, predicting both landmark coordinates and facial bounding boxes. The MDM [20] utilizes recurrent neural networks (RNN) for evenly distributing convolutional layers across cascade steps. RetinaFace [21] combines face detection, landmark localization, and detailed 3D face regression in a single framework. The SAN [22] uses style-aggregated images, enhancing stability against various image styles.

Heatmap regression models are based on the concept of creating heatmaps using landmark coordinates and then adjusting them [23], [24], [25], [26]. This method then converts the heatmap back into coordinates to identify the landmarks. LAB [23] introduced a boundary-aware technique using boundary lines for better alignment. AWing [24] introduced the adaptive Wing loss for optimizing the difference between foreground and background pixels. LUVLi [25] presented a loss that aids a CNN in simultaneously covering the uncertainty and visibility of landmarks.

Recent advancements show the Transformer model [27] surpassing conventional DNN, especially in the natural language processing. The Vision Transformer (ViT) [28] in the image domain, with its adaptive global attention, has become notable for tasks like object detection [29] and human pose estimation [30]. While still in the early stages in the domain of facial keypoints, pioneering works such as the SLPT [31] are beginning to surface. SLPT discerns the interrelation between landmarks through the Transformer and utilizing sparse local patches, it aligns them using a coarse-to-fine strategy.

### B. HEAD POSE ESTIMATION

In our study, accurate head pose estimation is essential for enhancing the precision of our remote pupil localization in A-pillar mounted cameras. This methodology adaptively adjusts the initial keypoint set, ensuring consistent and

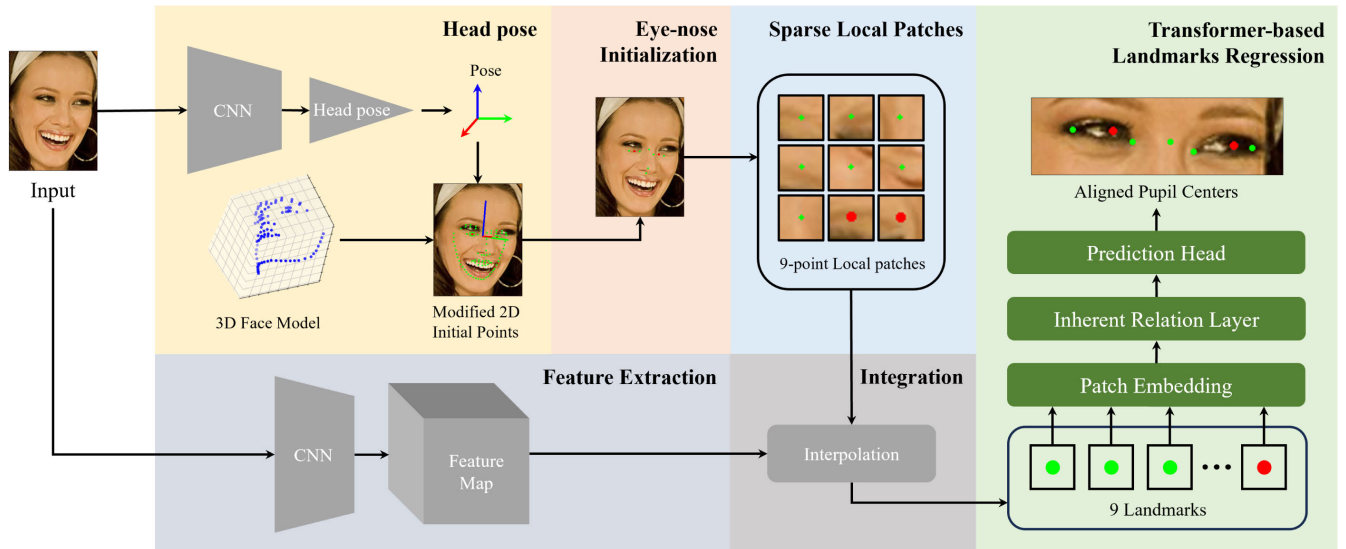
precise pupil tracking across a range of driving scenarios and head poses. The estimation of the head pose is widely used in areas like human motion [32], gaze estimation [33], and user attention recognition [34]. Classical methods for head pose estimation rely on appearance-based template models. These models compare test images with a set of pose samples to identify the best match for the head pose [35], [36]. This approach utilizes templates derived from face detectors, with each detector being specifically trained to recognize a particular head pose. As research continued, facial landmark detector-based template methods advanced significantly [37].

The growth of DCNN has brought in new methods for head pose estimation. Using shallow CNN and calculating regression loss to effectively estimate the head pose was proposed by [38]. Going a different way, KEPLER [39] used an iterative method based on a new Heatmap-CNN architecture, aimed at accurate 3D pose prediction. HyperFace [40] came out as a multi-task learning solution, built on a deep CNN foundation. This algorithm can detect faces, find landmarks, estimate poses, and by combining intermediate layers, it can even recognize gender. Later, the [41] was introduced as another multi-tasking learning algorithm. It could handle face detection, landmarks localization, pose estimation, gender recognition, smile detection, age estimation, and face identification and verification, all in one system. Taking a detailed approach, [42] used a multi-loss CNN. It focused on joining binned pose classification and Euler angle regression, offering a smart algorithm for head pose prediction. Recent improvements like MNN [43] combine head pose estimation within a face keypoints alignment model. Made with an encoder-decoder design, this model has a bottleneck structure and adds a head pose task at the end of the encoder to predict head pose.

### C. REMOTE PUPIL LOCALIZATION

Compared to facial landmark alignment techniques, there are not many studies that focus solely on remote pupil localization. Previous studies on remote pupil localization use frame-based RGB or NIR cameras to find the entire face area, eye-nose region, or eye region, and then extract the center of the pupil by regressing landmark points in the found region of interest. In autostereoscopic 3D displays, it is essential to accurately find the location of the pupil to separate the left and right stereoscopic images for the user [44]. A content-aware eye tracking method is proposed in [10]. This method distinguishes between contents such as thick eyeglasses, small eyes, and reflection of eyeglasses in the face image captured from an RGB camera, and performs SDM-based eye-nose points regression for each content. In the HCI field, commercial NIR camera-based methods are mainly used [45], which generate pupil cornea reflection and bright pupil using NIR light source, and then find the center of the pupil by simple image processing [46]. This method





**FIGURE 2.** Overview of the proposed head pose-aware regression method for pupil localization from A-pillar cameras. The process begins with a CNN-based backbone for 3D head pose estimation. Following this, optimized eye-nose initial points are derived for pupil localization. Sparse local patches around these eye-nose points are subsequently generated. These patches are then integrated with a CNN feature map and fed into the Transformer regression network, resulting in the final pupil center localization. The example face image is sourced from the WFLW dataset [23].

requires special hardware with camera sensors and NIR light sources located in specific locations.

Remote pupil tracking technology is also being actively researched for vehicular systems. In outdoor vehicular environments, driver facial images show many challenges such as diverse illumination conditions from oversaturation to low-light, occlusions due to various head poses and the use of sunglasses or hats, and the limited computing resources in vehicular embedded systems. An adaptable eye tracking technique that functions effectively in real-time, even with constrained system resources, was presented in [3]. Their approach combines classical machine learning methods like SDM [18] with CNN-based techniques such as practical facial landmark detection (PFLD) [47]. This combination effectively serves drivers with both bare faces and those wearing sunglasses. In DMS, lightweight traditional methods like Kalman filtering and support vector machine (SVM) are favored to run user monitoring algorithms based on eye tracking technology [48], [49]. A different approach presented in [50] assesses driver fatigue by monitoring yawning, blinking, and the duration of eye closure. It relies on the TCDCN [15] for face detection and pinpointing the center of the pupil, subsequently determining the driver's current driving state. However, many of these existing techniques rely solely on front cameras in vehicles and overlook the potential of the A-pillar camera.

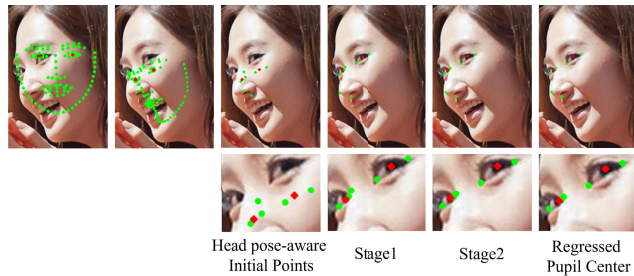
In summary, while existing remote pupil localization methods have made significant progress in parallel advancements in facial keypoints alignment techniques, they function optimally when the camera is positioned at a precise location for remote pupil tracking. However, these methods haven't been thoroughly studied for use with A-pillar cameras in

vehicles, where varied head poses pose unique challenges for remote eye tracking. To overcome these limitations, we introduce a novel approach: a head pose-aware pupil localization method. This method combines head pose estimation information with Transformer-based facial keypoints alignment technology, specifically designed to handle the diverse head movements of drivers when using the A-pillar camera. This approach allows for more accurate and reliable pupil tracking, even in challenging conditions such as varied head poses, changing light conditions, and partial occlusions commonly encountered in vehicular environments in AR 3D HUDs and DMS.

### III. METHOD

Our primary objective is to design a method that can efficiently localize the center of pupils using a camera positioned at the A-pillar, designed for advanced driver assistance systems in commercial vehicles. The challenge posed by the A-pillar's side location, as opposed to a front-facing position, means that the system has to contend with non-frontal views of the driver's face, encountering a variety of poses. This requires a pupil center tracking system that remains robust across these diverse orientations.

To address this challenge, we've expanded upon the traditional Transformer-based landmark regression method by incorporating the user's head pose data, ensuring consistent performance across diverse facial orientations. Our methodology consists of the following key steps. 1) Initialization of head pose-aware eye-nose keypoints: first, we generate keypoints sensitive to the head pose. This is achieved using a CNN-based head pose estimation network called the multi-task network (MTN) [51]. After this, we extract a set of nine



**FIGURE 3.** Illustration of the 6 steps in our proposed head pose-aware regression method for pupil localization from A-pillar cameras on WFLW dataset [23]. The steps are presented from left to right: Initial full facial keypoints, adjustments based on head pose estimation results, optimized eye-nose initial points for pupil localization, followed by regression stage 1, stage 2, and concluding with the final pupil center points.

optimized eye-nose points specifically designed for accurate pupil alignment. 2) Integration with localized patches and CNN-based features: once the nine head pose-aware eye-nose keypoints are identified, we create localized sparse patches around them. Features extracted from the original input image, using the methodology outlined in [52], are then combined with these patches. 3) Landmark regression using a Transformer: the created sparse landmark patches are input to the Transformer-based landmark regression approach, as described in [31]. This step finalizes the localization of the pupil center. A detailed visual representation of our proposed method is available in Figure 2.

#### A. HEAD POSE-AWARE TRANSFORMER-BASED PUPIL REGRESSION

Our proposed head pose-aware pupil regression method is built upon two foundational studies. Firstly, the CNN-based head pose estimation algorithm, MTN [51], and secondly, the Transformer-based facial landmark localization network, sparse local patch Transformer, SLPT [31]. While our approach is grounded in the SLPT keypoints regression technique, it differs from the original by adopting a head pose-aware approach. Given the unique perspective challenges of a camera positioned at the A-pillar, we enhanced the SLPT by integrating a head pose module. For the head pose estimation task, we utilized MTN [51].

SLPT [31] is a Transformer-based network designed for facial keypoints alignment, utilizing on the inherent relationship between facial keypoints. It begins its process by creating sparse local patches from initial facial keypoints derived from an average shape of facial datasets. These patches, once combined with a pre-processed CNN-based feature map, enter the Transformer network. Unlike the traditional Vision Transformer (ViT) [28], which breaks down the entire image into multiple patches for the attention mechanism, SLPT focuses solely on patches centralized around landmarks. The SLPT Transformer network accepts these sparse local patches and progresses through patch embedding, an attention mechanism termed as the inherent relation layer, and finally passes through a prediction head,

which uses multi-layer perceptron. The outcome is a set of regressed optimal shape points. This hybrid model utilizes the local feature extraction capability of CNNs and the global feature capturing attribute of Transformers. Furthermore, with each iteration through the Transformer network, the sparse local patches are cropped based on the landmarks predicted in the previous stage, gradually converging to the optimal facial landmarks in a coarse-to-fine regression framework.

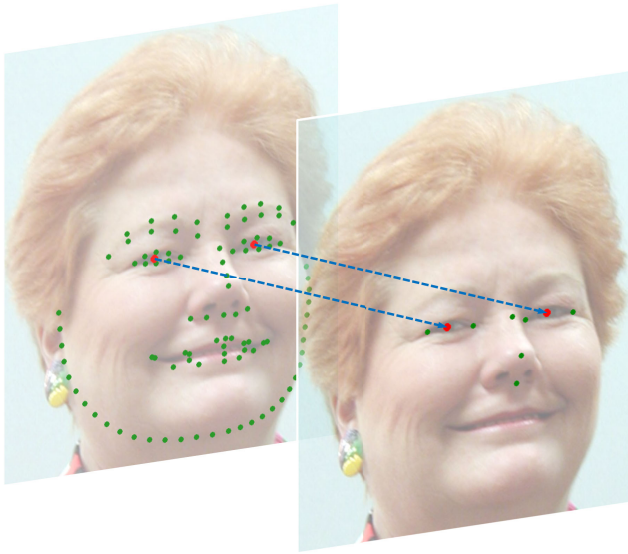
To address the specific needs of an A-pillar camera in vehicles, we adapted the Transformer-based SLPT [31] algorithm into a head pose-aware format. In constructing our head pose-aware network, a head pose module was incorporated into the SLPT. The emphasis for this enhancement was the starting point of regression, the mean face initial point. While conventional facial regression networks utilize a traditional two-dimensional (2D) mean face initial point, our approach integrates it with head pose estimation, resulting in a head pose-aware initial point. The head pose estimation network uses the multi-task network (MTN) [51]. MTN is a head pose estimation approach that utilizes a stacked hourglass (HG) [53] encoder-decoder network architecture. Within this architecture, an additional head pose encoder is connected to the bottleneck of the encoder, enabling the estimation of a 3D head pose. Once this 3D head pose is obtained, it is used to rotate the 3D face model, specifically AFLW [54]. After that, a 2D projection is done, which results in the creation of our head pose-aware initial points. In the subsequent stages of our algorithm, these initial points are refined into nine optimized eye-nose initial points. Only these selected points are then input into the SLPT [31] regression module. This method leads to the precise determination of the regressed pupil center. A step-by-step breakdown of this regression process can be illustrated in Figure 3.

#### B. 9-POINT ALIGNMENT FOR PUPIL LOCALIZATION FROM A-PILLAR CAMERAS

Traditional facial keypoints alignment algorithms predominantly target the full facial region, relying on comprehensive keypoints such as 98 points [23] and 68 points [55], [56]. However, our primary objective is the precise localization of the pupil center. Thus, instead of looking at all facial keypoints, we selectively concentrated on essential facial points specific to our study. Utilizing the full set of facial keypoints for pupil localization can cause inaccuracies. Specifically, the algorithm learns to minimize the error loss across all keypoints. Consequently, under dynamic illumination conditions encountered in vehicular settings, or due to facial motions, inaccuracies in the regression of points such as face boundary points or mouth points can increase alignment errors in the pupil center. For instance, when using A-pillar cameras in vehicles, the challenges are manifold: occlusions from hands on the steering wheel, wearables like caps causing shadowing, or even dynamic movements of the mouth, which are unrelated to pupil movements. To focus exclusively on

**TABLE 1.** Performance evaluation of our proposed model across various WFLW [23] test subsets, measured using NME, FR<sub>10</sub>, and AUC<sub>10</sub> metrics.

Metric	All	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%) (↓)	2.79	5.28	2.84	2.76	2.61	3.48	3.53
FR <sub>10</sub> (%) (↓)	1.28	6.14	1.59	1.00	1.46	2.85	2.20
AUC <sub>10</sub> (↑)	0.731	0.524	0.735	0.732	0.743	0.673	0.671

**FIGURE 4.** Example images (WFLW dataset [23]) with 9-point alignment for pupil localization. Instead of the conventional full facial keypoints commonly used in public face databases, we adopt a set of points optimized specifically for pupil center localization in our Transformer-based regression network. This consists of 9 selected points, excluding those such as face boundary and mouth landmarks that could decrease the precision of pupil localization regression.

the pupil center localization, we carefully selected keypoints robust against varied illumination, occlusion due to driving, and facial poses via an A-pillar camera.

Through extensive testing with diverse facial images captured from the A-pillar, we identified an optimal set of nine points suited to our objectives. While drivers mainly look ahead, we considered head rotations of up to approximately 30 degrees to either side to ensure clear visibility while driving. Given this angular assumption, we utilized our keypoint selection to ensure the visibility of both the endpoints of the eyes. As a result, we opted for keypoints as follows: the medial canthus, lateral canthus, and pupil center from the eyes, bridge of the nose, nose tip, and subnasale. Figure 4 visualizes these nine selected points.

In our proposed head pose-aware Transformer-based pupil regression algorithm, these keypoints are used during the head pose-aware initial points phase. As shown in the first and second steps of Figure 3, our algorithm first utilizes the full facial keypoints, for example, the 98 points in WIDER Facial Landmarks in-the-wild (WFLW) [23], for user head pose estimation. After determining the head pose, our system transitions to the Transformer-based regression phase, avoiding the full facial keypoints and adopting the

**FIGURE 5.** Visualization of our proposed method for pupil center localization using the public face dataset, WFLW [23]. Red dots represent the pupil centers, while the green dots illustrate the remaining 7 eye-nose landmarks.

specialized 9-point scheme introduced in this paper (3rd step in Figure 3). Rather than using all facial landmarks, our method generates sparse local patches from these 9 eye-nose keypoints. This refined input is then processed through the Transformer-based landmark regression to precisely adjust these 9 keypoints (4th to 6th steps in Figure 3).

#### IV. EXPERIMENTAL RESULTS

In this research, we evaluated our proposed head pose-aware pupil localization algorithm on the popular public dataset, WFLW [23], as well as our in-house dataset captured from A-pillar cameras mounted in real vehicles.

##### A. DATASETS

In the absence of a dedicated benchmark dataset for remote pupil localization, we have chosen the WFLW dataset [23] as a suitable alternative for our evaluations. Recognized for its comprehensive and challenging nature in the facial keypoints alignment domain, the WFLW dataset provides a robust platform to assess the effectiveness of our proposed pupil localization method across a variety of driving conditions. It is comprised of facial images taken from wild



**TABLE 2.** Performance evaluation of our proposed model across various in-house test subsets, measured using NME, FR<sub>10</sub>, and AUC<sub>10</sub> metrics.

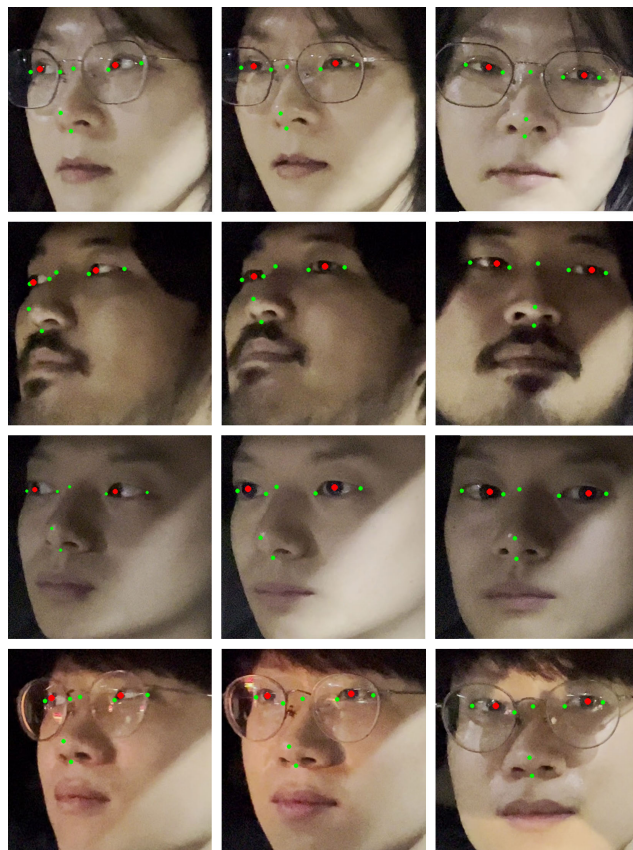
Metric	All	Front	Side
NME(%)	2.96	2.67	3.32
FR <sub>10</sub> (%)	0.72	0.00	1.61
AUC <sub>10</sub>	0.707	0.733	0.674

environments. It contains 7,500 training images and 2,500 test images, each annotated with 98 keypoints. The dataset is categorized based on six attributes: pose, expression, illumination, make-up, occlusion, and blur. Additionally, it's grouped into 62 diverse categories such as parades and sports. Particularly, it incorporates a significant amount of challenging pose data. Thus, it's suited for training and testing our A-pillar camera-based head pose-aware pupil localization algorithm. Unlike other famous datasets like 300W [56], which lacks pupil center localization annotation, the 98 annotations of WFLW include the pupil center, making it an ideal primary experimental dataset for our purposes.

To enhance the validation of our algorithm, we not only relied on public datasets but also created an in-house dataset. This dataset was generated from cameras mounted on the A-pillar of vehicles, capturing diverse head poses of real drivers. The recordings were made with six volunteers simulating driving scenarios, which were captured in a resolution of  $1920 \times 1080$ . For the application of our head pose-aware localization algorithm, we processed these images by manually cropping only the facial regions and subsequently resizing them to  $256 \times 256$  dimensions for optimal input. It's important to note that the distance between the driver and the A-pillar-mounted camera was approximately 0.5 meters. Abiding by our study's premise, which assumes that drivers mainly face forward but might rotate their heads up to 30 degrees on either side for better visibility, we curated a final test dataset of 100 images. Our tests were specifically limited to head poses where both lateral canthus were clearly visible. These 139 test images underwent manual nine eye-nose keypoint annotation by our team. Of these, 77 represent drivers looking straight ahead, while 62 capture slight lateral head movements within around the 30-degree range. All human participants in this study provided their informed consent.

## B. EVALUATION METRICS

For our head pupil center localization task, we adopted standard metrics widely used in facial keypoint localization studies. A key metric for accuracy is the Normalized Mean Error (NME), for which we adopt a normalization process inspired by [23]. In this statistic, the distance between the outer corners of the eyes termed the "inter-ocular" distance, is used as the normalization factor. In addition to the NME, we also utilized two more metrics: the Failure Rate (FR) for sufficiency and the Area Under the Curve (AUC) for reliability. The FR represents the proportion of test images

**FIGURE 6.** Visualization of our proposed method for pupil center localization using an in-house dataset. Images show various poses captured inside a vehicle with a camera positioned at the A-pillar. Red dots denote the pupil centers, while green dots highlight the remaining 7 eye-nose landmarks.

where the NME exceeds a given threshold. On the other hand, the AUC provides a measure of the area under the Cumulative Error Distribution (CED) curve. In this study, we standardized the thresholds for both FR and AUC to 10%.

## C. IMPLEMENTATION DETAILS

For our pupil localization task, we designed our network model utilizing the PyTorch framework and executed the computations on an Ubuntu 20.04.6 LTS system equipped with an RTX 4090 (24GB) GPU. All training and testing images were manually cropped based on facial bounding boxes and resized to a uniform resolution of  $256 \times 256$ . Our pupil regression algorithm uses the adaptive head pose-aware initial keypoints for each individual face. These specially modified head pose initial points were derived using a pre-trained model from MTN [51] head pose estimation. Once these head pose-informed initial points were obtained, they were combined with the cropped facial image to train the SLPT [31] regression model. For the CNN-based feature extraction within the SLPT model's architecture, we used HRNetW18C [52] as the backbone network. The training process maintained hyperparameters in the original SLPT

**TABLE 3.** Performance comparison highlighting the effects of head pose and number of points used in WFLW [23] dataset. Metrics include NME, FR<sub>10</sub>, and AUC<sub>10</sub>. Performance was calculated for all the keypoints alignment.

Method	NME(%)	FR <sub>10</sub> (%)	AUC <sub>10</sub>
Full Facial 98 points [31]	4.14	2.76	0.595
Full Facial 98 points + Head Pose	4.09	2.48	0.602
<b>Eye-nose 9 points + Head Pose (Ours)</b>	<b>2.79</b>	<b>1.28</b>	<b>0.731</b>

**TABLE 4.** Performance comparison highlighting the effects of head pose and number of points used in our custom in-house dataset. Metrics include NME, FR<sub>10</sub>, and AUC<sub>10</sub>. Performance was calculated for only pupil center alignment.

Method	NME(%)	FR <sub>10</sub> (%)	AUC <sub>10</sub>
Full Facial 98 points [31]	1.56	0.00	0.844
Full Facial 98 points + Head Pose	1.53	0.00	0.847
<b>Eye-nose 9 points + Head Pose (Ours)</b>	<b>1.31</b>	<b>0.00</b>	<b>0.869</b>

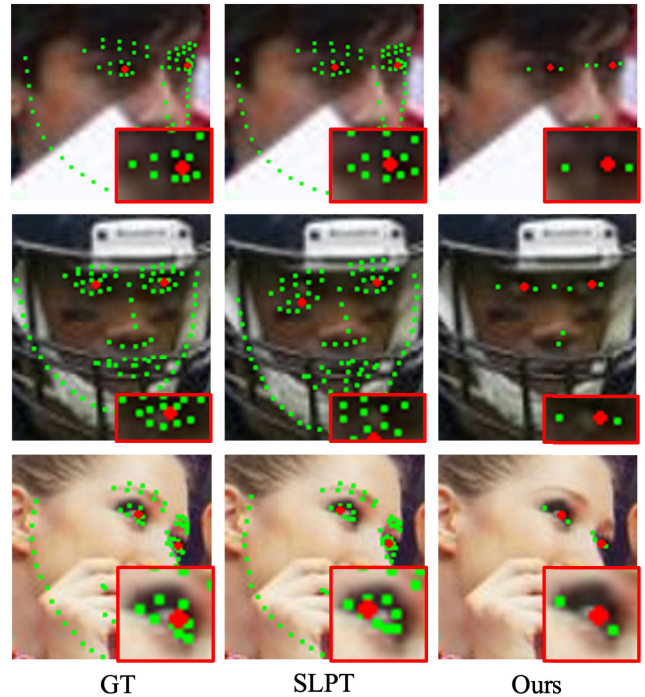
implementation, using a batch size of 32 and a learning rate of 0.001.

**D. EVALUATION ON WFLW**

We conducted an evaluation of our proposed method using the publicly available WFLW dataset [23]. In our experiments, we utilized 2,500 images from the WFLW dataset for testing. Across all test datasets, our method achieved an NME of 2.79%, an FR<sub>10</sub> of 1.28%, and an AUC<sub>10</sub> of 0.731. In the pose subset of the WFLW, which includes a diverse range of facial poses, the NME slightly increased to 5.28%, indicating a more challenging subset compared to the entire dataset. Detailed results from the evaluation across the six attribute subsets of the WFLW dataset can be found in Table 1. Figure 5 provides visual examples from the WFLW test dataset, especially focusing on challenging images featuring distinct poses and occlusions, further demonstrating the effectiveness of our proposed method.

**E. EVALUATION ON IN-HOUSE DATASET**

To further validate the robustness of our proposed algorithm, we evaluated its performance on a set of 139 images from our in-house dataset. Across this dataset, the proposed pupil localization method yielded an NME of 2.96%, an FR<sub>10</sub> of 0.72%, and an AUC<sub>10</sub> of 0.707. Upon analyzing the in-house dataset results, we observed an NME of 2.67% in the front subset and 3.32% in the side subset. Interestingly, due to the distinct positioning of the A-pillar mounted camera, the front subset, where the driver mainly faces forward, showed a slightly higher NME. Detailed outcomes from this in-house



**FIGURE 7.** Comparison of the ground truth, face alignment results of the original SLPT [31], and our proposed regression method on faces with various poses, blur, make-up, and occlusion from WFLW [23]. Detected pupil centers are emphasized and enlarged in red boxes for clarity.

evaluation are tabulated in Table 2. Figure 6 shows a variety of results from both the front and side subsets of our in-house dataset, further demonstrating the algorithm’s performance across different orientations.

**V. DISCUSSION**

Our proposed method shows outstanding performance on both the challenging and widely recognized WFLW [23] public face dataset and our custom in-house dataset, which captures driver faces directly from an A-pillar camera. Specifically, the algorithm achieves an NME of 2.79%, FR of 1.2%, and AUC of 0.731 on the entire WFLW test dataset. Furthermore, to verify the effectiveness and potential of our head pose-aware pupil localization, we tested it on various subsets of WFLW, including the pose attribute subset. Even in such challenging subsets, the algorithm still demonstrated impressive performance with an NME of 5.28%, FR of 6.4%, and AUC of 0.524, despite a slight decline compared to the complete dataset results.

For a more realistic evaluation, in addition to the public WFLW dataset, we also assessed our algorithm’s performance on our in-house dataset captured from an A-pillar mounted camera in vehicles. This dataset is divided into front and side subsets, specifically capturing the varied head poses typical in driving scenarios. On this dataset, the algorithm achieves an NME of 2.96% on the entire dataset, 2.67% on the front subset, and 3.32% on the side subset. Interestingly, while the in-house dataset registers a slightly higher NME

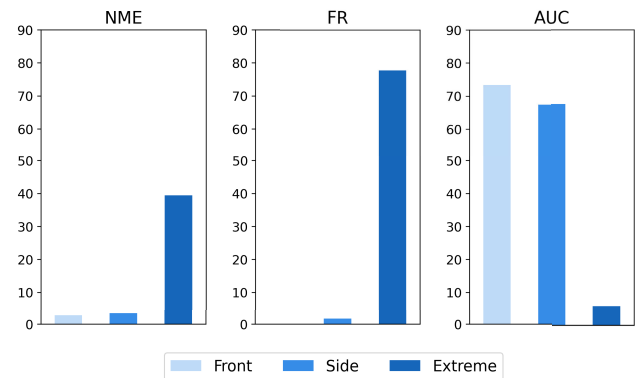


of 2.96% compared to the 2.79% of the WFLW [23] full test dataset, this can be attributed to the inherently greater pose variations in the in-house dataset due to the unique positioning of the A-pillar mounted camera. The front subset of our custom in-house dataset outperforms the pose subset of WFLW, achieving an NME of 2.67% compared to 5.28%. This can be attributed to our experimental setup in the in-house dataset, which ensures clear visibility of both eyes and excludes extreme head poses such as 90-degree rotations. Notably, in our in-house dataset, the FR demonstrates excellent results, with a remarkable 0% in the front subset, which aligns with our study's premise that drivers primarily face forward. In the side subset, where head rotation of up to 30 degrees is considered, the FR is still low at 1.61%, indicating sufficient and robust performance even under moderate head pose variations. Additionally, considering the 3D margin of AR 3D HUDs [3], our results demonstrate high accuracy for commercial applications. Furthermore, our method's AUC of 0.707, a comprehensive indicator of overall reliability, highlights the reliability and robustness of our approach in diverse driving conditions.

#### Effects of Head Pose and 9-Point Alignment:

Our proposed method integrates two primary elements: head pose estimation and an optimized set of eye-nose keypoints for precise pupil localization. We validated the effectiveness of our method on the WFLW dataset [23]. Given that our pupil regression algorithm is derived from the SLPT [31], which utilizes a full set of 98 facial keypoints, we conducted an ablation study using this model as a foundation. We analyzed the influence of our primary components on established metrics such as NME, FR, and AUC, with detailed results presented in Table 3 and Table 4. Performance metrics in Table 3 were determined for all the keypoint alignments, whereas those in Table 4 focused solely on the accuracy of pupil center points.

It's important to emphasize that head pose estimation is an integral component of our algorithm. As shown in Table 3, integrating pose estimation with the original SLPT [31] led to a decrease in NME from 4.14% to 4.09%. Additionally, we observed a reduction in FR by 0.28% and an improvement in AUC by 0.007, marking overall performance enhancements. Also shown in Table 4, which presents the evaluation of our in-house data, indicates that integrating pose estimation with the original SLPT led to a decrease in NME from 1.56% to 1.53%. Furthermore, we noted an improvement in AUC by 0.003, signifying overall performance enhancements. Subsequently, we evaluated the efficacy of our proposed 9-point scheme for pupil localization. When compared to the 98-point facial recognition method enhanced with head pose estimation, our 9-point approach resulted in a decrease in NME from 4.09% to 2.79%, a reduction in FR from 2.48% to 1.28%, and an increase in AUC from 0.602 to 0.731. Similarly, for another dataset, the 9-point method brought the NME down from 1.53% to 1.31% and increased the AUC from 0.847 to 0.869. This analysis clearly demonstrates that



**FIGURE 8. Comparative analysis of performance metrics in various head poses. This figure illustrates the NME, FR<sub>10</sub>, and AUC<sub>10</sub> for our remote pupil localization method. The metrics are compared across different in-house test subsets, including moderate head poses (front, side) and extreme poses.**

both the head pose integration and our customized 9-point configuration enhance the original SLPT's performance. Figure 7 provides a visual representation of our findings on the WFLW dataset, contrasting the ground truth, the original SLPT, and our pupil localization technique, thereby highlighting the precision of our detection method. To further investigate the performance of our method in extreme head poses, we conducted additional experiments and presented the results in Figure 8. This figure illustrates the performance of our method in scenarios with head poses exceeding 60 degrees. As shown, while our method performs well in moderate head pose variations, there is a noticeable performance drop in scenarios with extreme head poses, indicating an area for future enhancements.

Our approach improves upon the original SLPT by recognizing the user's head pose and modified initial points closely to optimal locations using head pose information, enhancing regression accuracy. Additionally, by eliminating unnecessary points and focusing solely on essential keypoints for pupil regression, our method minimizes errors caused by irrelevant points, resulting in superior quantitative performance.

#### A. COMPARISON WITH EXISTING APPROACHES

To demonstrate the competence of our algorithm, we conducted a comparison with the existing state-of-the-art facial keypoint regression algorithms. Our evaluations, particularly on the public WFLW dataset [23], are presented in Table 5. Impressively, our method outperforms previous approaches across all performance metrics, including NME, FR, and AUC. It's worth emphasizing that our method registered an NME of 2.79% on the full WFLW dataset, surpassing the most competitive state-of-the-art method, ADNet [26], which achieved 4.14%. Furthermore, when evaluating the WFLW pose subsets, our algorithm stands out distinctly. The best performance from existing methods recorded an NME of 6.96%, whereas ours achieved 5.28%. Our algorithm,

**TABLE 5. Comparing with state-of-the-art methods on WFLW [23]. The best results are marked in bold. Metrics include NME, FR<sub>10</sub>, and AUC<sub>10</sub>.**

Dataset	Method	NME(%)	FR <sub>10</sub> (%)	AUC <sub>10</sub>
Full	LAB [23]	5.27	7.56	0.532
	AVS+SAN [57]	4.39	4.08	0.591
	LUVLi [25]	4.37	3.12	0.557
	AWing [24]	4.36	2.84	0.572
	ADNet [26]	4.14	2.72	0.602
	SLPT [31]	4.14	2.76	0.595
	<b>Ours</b>	<b>2.79</b>	<b>1.28</b>	<b>0.731</b>
Pose	LAB [23]	10.24	28.83	0.235
	AVS+SAN [57]	8.42	18.10	0.311
	LUVLi [25]	7.56	15.95	-
	AWing [24]	7.21	9.20	0.334
	ADNet [26]	6.96	12.72	0.344
	SLPT [31]	6.96	12.27	0.348
	<b>Ours</b>	<b>5.28</b>	<b>6.14</b>	<b>0.524</b>

**TABLE 6. Computational complexity of our method and state-of-the-art methods.**

Method	FLOPs (G)	Time (ms)
LAB [23]	18.85	60
AVS+SAN [57]	33.87	-
AWing [24]	26.8	29
ADNet [26]	17.04	95.29
SLPT [31]	6.12	9
<b>Ours</b>	<b>20.82</b>	<b>18.3</b>

similar to other state-of-the-art methods, exhibited decreased performance on the pose subset compared to the full dataset. This reduction in performance was due to the pose subset predominantly comprising large-angle head poses, which inherently presented greater challenges for accurate pupil localization. A comprehensive comparison between our novel algorithm and state-of-the-art deep learning techniques, specifically LAB [23], SAN [22], LUVLi [25], AWing [24], ADNet [26], and SLPT [31], is provided in Table 5. Among heatmap regression methods, ADNet [26] showed performance nearly comparable to the Transformer-based method, SLPT [31]. However, by integrating head pose estimation information and our optimal eye-nose 9 points, our results surpass both ADNet and SLPT. In addition to performance metrics, we have also addressed computational complexity, as detailed in Table 6. Our algorithm, while exhibiting higher computational complexity than some state-of-the-art methods, demonstrates a necessary trade-off to achieve higher accuracy and robustness, especially in A-pillar mounted camera scenarios. This balance of complexity and effectiveness aligns with [58], where increased system sophistication leads to improved adaptability and precision in dynamic environments.

Differentiating from facial landmark point alignment technologies, studies exclusively dealing with remote eye tracking are relatively sparse. They either do not offer

precision in pupil center localization, or they use varying testing datasets, making direct comparisons on public datasets, similar to those used for facial keypoint alignment, practically impossible. It's also notable that many of these prior works rely on front cameras instead of A-pillar mounted ones. For instance, the study in [6] developed a driver fatigue monitoring system based on the TCDCN method, using a front camera. While they achieved a 95% drowsiness detection rate using their custom in-house dataset, they did not provide the precision of pupil center localization. Their lack of reliance on the A-pillar camera suggests they did not account for diverse head poses. Our primary reference for remote pupil localization, [3], utilized two distinct methods for scenarios with and without sunglasses, ensuring coverage for occlusions. Their evaluation spanned both their in-house dataset and the public WFLW dataset. On the WFLW, their sunglasses tracker registered a 7.43% NME while their bare face tracker achieved 1.71% NME. Although their bare face tracker, tested on a selected subset of WFLW where the pupils were clearly visible, outperformed our method, a direct comparison is complicated since our evaluation includes the entire WFLW dataset [23], covering challenging cases with various poses.

## VI. CONCLUSION

In this paper, we have effectively integrated a head pose-aware regression approach into traditional pupil localization methods that primarily used facial keypoints alignment. By understanding the relationship between a person's pupil position and head pose, our method ensures consistent pupil center localization, even in different head poses. Furthermore, with our proposed set of nine eye-nose points, we've improved the accuracy of pupil detection, particularly when other facial features might be hidden. Our method has demonstrated strong performance on both the well-known public WFLW dataset and our specific in-house dataset, emphasizing its potential for commercial eye tracking using A-pillar mounted cameras.

While our proposed method has yielded impressive results, there were a few limitations in our study. One of the limitations is the relatively small size of our in-house test dataset, primarily due to the challenges associated with manual keypoint labeling. Additionally, although the in-house dataset was captured using A-pillar mounted cameras in vehicles, it lacks diversity in terms of driving scenarios. Conditions like low light environments, backlighting, varying illumination levels, and occlusions were not extensively covered. Creating a comprehensive dataset that includes these scenarios is challenging, not only because of the inherent risks associated with ensuring volunteer driving safety but also due to the costs of manual labeling. Moreover, our method currently struggles with extreme head poses, particularly those exceeding 60 degrees, such as when the driver looks towards the right side rear mirror at an angle of 90 degrees. Addressing these extreme head poses presents a significant opportunity for future research. Expanding our

approach to integrate occlusion handling techniques with our head pose-aware strategy could offer a more comprehensive solution, enhancing the adaptability and effectiveness of our method for a range of vehicular applications. This direction is essential for developing a robust commercial solution that can effectively handle significant head poses with A-pillar mounted cameras. Furthermore, a deep study on optimizing computational efficiency while retaining high performance using pyramid interconnection networks [59] may provide a more efficient study for remote pupil tracking.

## REFERENCES

- [1] Z. An, X. Xu, J. Yang, Y. Liu, and Y. Yan, "A real-time three-dimensional tracking and registration method in the AR-HUD system," *IEEE Access*, vol. 6, pp. 43749–43757, 2018.
- [2] H. Lee, J. Lee, and S. Hong, "Crosstalk reduction method in a glasses-free AR 3D HUD," *Proc. SPIE*, vol. 11765, Mar. 2021, Art. no. 1176516.
- [3] D. Kang and L. Ma, "Real-time eye tracking for bare and sunglasses-wearing faces for augmented reality 3D head-up displays," *IEEE Access*, vol. 9, pp. 125508–125522, 2021.
- [4] P. KIELTY, M. S. Dilmaghani, W. Shariff, C. Ryan, J. Lemley, and P. Corcoran, "Neuromorphic driver monitoring systems: A proof-of-concept for yawn detection and seatbelt state detection using an event camera," *IEEE Access*, vol. 11, pp. 96363–96373, 2023.
- [5] C. Ryan, A. Elrasad, W. Shariff, J. Lemley, P. KIELTY, P. Hurney, and P. Corcoran, "Real-time multi-task facial analytics with event cameras," *IEEE Access*, vol. 11, pp. 76964–76976, 2023.
- [6] J. Blattgerste, P. Renner, and T. Pfeiffer, "Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views," in *Proc. Workshop Commun. Gaze Interact.*, Jun. 2018, pp. 1–9.
- [7] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Dec. 2016.
- [8] J. Wang, S. Xu, Y. Dai, and S. Gao, "An eye tracking and brain-computer interface based human-environment interactive system for amyotrophic lateral sclerosis patients," *IEEE Sensors J.*, vol. 23, no. 20, pp. 24095–24106, Oct. 2023.
- [9] Z. Zhu and Q. Ji, "Novel eye gaze tracking techniques under natural head movement," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2246–2260, Dec. 2007.
- [10] D. Kang and J. Heo, "Content-aware eye tracking for autostereoscopic 3D display," *Sensors*, vol. 20, no. 17, p. 4787, Aug. 2020.
- [11] Z. Jankó and L. Hajder, "Improving human-computer interaction by gaze tracking," in *Proc. IEEE 3rd Int. Conf. Cognit. Infocommun.*, Kosice, Slovakia, Dec. 2012, pp. 155–160.
- [12] G. Chen, L. Hong, J. Dong, P. Liu, J. Conradt, and A. Knoll, "EDDD: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor," *IEEE Sensors J.*, vol. 20, no. 11, pp. 6170–6181, Jun. 2020.
- [13] L. G. T. Ribas, M. P. Cocron, J. L. Da Silva, A. Zimmer, and T. Brandmeier, "In-cabin vehicle synthetic data to test deep learning based human pose estimation models," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 610–615.
- [14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Kauai, HI, USA, Dec. 2001, pp. 511–518.
- [15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [16] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, Jan. 2018.
- [17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
- [18] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 532–539.
- [19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [20] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4177–4187.
- [21] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5202–5211.
- [22] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 379–388.
- [23] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2129–2138.
- [24] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6970–6980.
- [25] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng, "LUVLi face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8233–8243.
- [26] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei, "ADNet: Leveraging error-bias towards normal direction in face alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3060–3070.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Dec. 2017, pp. 5998–6008.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2021, pp. 1–12.
- [29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, Aug. 2020, pp. 213–229.
- [30] S. Yang, Z. Quan, M. Nie, and W. Yang, "TransPose: Keypoint localization via transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11782–11792.
- [31] J. Xia, W. Qu, W. Huang, J. Zhang, X. Wang, and M. Xu, "Sparse local patch transformer for robust face alignment and landmarks inherent relation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 4042–4051.
- [32] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, Nov. 2006.
- [33] R. Ranjan, S. De Mello, and J. Kautz, "Light-weight head pose invariant gaze tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2237–22378.
- [34] T. Singh, M. Mohadikar, S. Gite, S. Patil, B. Pradhan, and A. Alamri, "Attention span prediction using head-pose estimation with deep neural networks," *IEEE Access*, vol. 9, pp. 142632–142643, 2021.
- [35] J. Ng and S. Gong, "Composite support vector machines for detection of faces across views and pose estimation," *Image Vis. Comput.*, vol. 20, nos. 5–6, pp. 359–368, Apr. 2002.
- [36] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: An application to face detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* San Juan, Puerto Rico, Jun. 1997, pp. 130–136.
- [37] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
- [38] M. Patachiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017.
- [39] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 258–265.



- [40] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
- [41] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 17–24.
- [42] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2155–215509.
- [43] R. Valle, J. M. Buenaposada, and L. Baumela, "Multi-task head pose estimation in-the-wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2874–2881, Aug. 2021.
- [44] S. Lee, J. Park, J. Heo, B. Kang, D. Kang, H. Hwang, J. Lee, Y. Choi, K. Choi, and D. Nam, "Autostereoscopic 3D display using directional subpixel rendering," *Opt. Exp.*, vol. 26, no. 16, p. 20233, Aug. 2018.
- [45] A. Gibaldi, M. Vanegas, P. J. Bex, and G. Maiello, "Evaluation of the Tobii EyeX eye tracking controller and MATLAB toolkit for research," *Behav. Res. Methods*, vol. 49, no. 3, pp. 923–946, Jun. 2017.
- [46] C. Mestre, J. Gautier, and J. Pujol, "Robust eye tracking based on multiple corneal reflections for clinical applications," *J. Biomed. Opt.*, vol. 23, no. 3, Mar. 2018, Art. no. 035001.
- [47] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "PFLD: A practical facial landmark detector," 2019, *arXiv:1902.10859*.
- [48] X. Liu, F. Xu, and K. Fujimura, "Real-time eye detection and tracking for driver observation under various light conditions," in *Proc. Intell. Vehicle Symp.*, Versailles, France, Jun. 2002, pp. 344–351.
- [49] J. Jo, "Vision-based method for detecting driver drowsiness and distraction in driver monitoring system," *Opt. Eng.*, vol. 50, no. 12, Dec. 2011, Art. no. 127202.
- [50] W. Deng and R. Wu, "Real-time driver-drowsiness detection system using facial features," *IEEE Access*, vol. 7, pp. 118727–118738, 2019.
- [51] A. Prados-Torreblanca, J. M. Buenaposada, and L. Baumela, "Shape preserving facial landmarks with graph attention networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, London, U.K., Nov. 2022, pp. 155.1–155.13.
- [52] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
- [53] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, Sep. 2016, pp. 483–499.
- [54] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.
- [55] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1513–1520.
- [56] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sydney, NSW, Australia, Dec. 2013, pp. 397–403.
- [57] S. Qian, K. Sun, W. Wu, C. Qian, and J. Jia, "Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 10152–10162.
- [58] M. H. Shojaeefard, M. Mollajafari, S. Ebrahimi-Nejad, and S. Tayebi, "Weather-aware fuzzy adaptive cruise control: Dynamic reference signal design," *Comput. Electr. Eng.*, vol. 110, Sep. 2023, Art. no. 108903.
- [59] H. S. Shahhoseini, E. S. Kandzi, and M. Mollajafari, "Nonflat surface level pyramid: A high connectivity multidimensional interconnection network," *J. Supercomput.*, vol. 67, no. 1, pp. 31–46, Jan. 2014.



**DONGHWA KANG** is currently pursuing the bachelor's degree with the School of Electronic and Electrical Engineering, Hongik University, Seoul, South Korea. Her research interests include the area of computer vision, including human pose estimation, face keypoints detection, and alignment.



**DONGWOO KANG** (Member, IEEE) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2009 and 2013, respectively. He was a Senior Researcher with the Samsung Advanced Institute of Technology, Suwon, South Korea, from 2013 to 2021.

In 2021, he joined the Faculty of the Department of Electronic and Electrical Engineering, Hongik University, Seoul, where he is currently an Assistant Professor. His research interests include the area of image processing and computer vision including detection, tracking, segmentation, image enhancement, application to augmented reality, autostereoscopic 3D displays, and medical image analysis.

• • •