IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 November 2023, accepted 11 December 2023, date of publication 14 December 2023,
date of current version 20 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3343152

RESEARCH ARTICLE

Event Camera-Based Pupil Localization: Facilitating Training With Event-Style Translation of RGB Faces

DAEHYUN 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 An innovative approach to pupil tracking using event cameras is presented in this paper. Our method incorporates two primary processes: RGB-to-Event image domain translation and pupil localization utilizing Event Cameras. Initially, we convert traditional RGB images into event-like images with our novel adaptive StyleFlow algorithm. This advanced algorithm enables the generation of images that are remarkably similar to those produced by real event cameras in terms of their distinctive characteristics and visual appeal. Subsequently, we perform the pupil localization process, which involves applying the RetinaFace algorithm. This algorithm is trained using our unique cross-modal learning strategy on a mixed dataset, consisting of both RGB and the newly transformed event-like images. When evaluated using real event camera data, our approach sets a new benchmark in accuracy performance. We achieved a face detection accuracy of 99.4% and a pupil alignment accuracy of 97.2%, exceeding the performance of previous deep learning-based methods that were trained on conventional RGB images. Our results effectively demonstrate the promising potential of event camera-based pupil tracking. Furthermore, our study represents an important advance in the field, offering the possibility of future advancements and potential applications in vehicular systems and augmented reality heads-up displays (AR HUDs).

INDEX TERMS Cross-modal learning, dynamic vision sensor, event cameras, event image generation, pupil detection, pupil localization, RGB-to-event image domain translation, training strategy.

I. INTRODUCTION

Eye tracking, which locates the center of the pupil and estimates gaze direction, is a core technology with diverse applications in various fields. It plays a significant role in attention tracking for market research and advertising [1], enhancing human-robot interaction [2], and enabling advanced features in automotive applications, including driver monitoring systems (DMS) [3] and head-up displays (HUDs) [4], [5]. Additionally, eye tracking finds utility in augmented reality (AR) [6], virtual reality (VR) [7], and three-dimensional (3D) display systems [8], as well as in consumer devices such as mobile smartphones and laptops [8], offering enhanced user interactions and gaming experiences. In market and advertising research, eye tracking is utilized to monitor consumers' attention and analyze the performance of products [1]. Regarding DMS in vehicular applications, eye tracking becomes increasingly important for detecting and monitoring driver status [3]. HUDs enable the generation of natural 3D content aligned with the user's eye position [4], [5]. For AR, VR, and Autostereoscopic 3D display applications, eye tracking is essential in reducing 3D fatigue by ensuring the accurate separation of left and right stereoscopic images, thus providing comfortable viewing experiences [8], [9]. Moreover, the integration of eye tracking consumer devices allows for various user interactions in various applications [2], [10]. Additionally, eye tracking is an essential tool for human behavior analysis, where rapid eye

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

FIGURE 1. Comparison examples between traditional frame-based RGB camera images (left) and event camera images (right). Event camera images demonstrate consistency compared to the RGB camera images under various light conditions: normal light (top), low light (middle), and image saturation in sunlight (bottom).

movement estimation helps in analyzing user interactions for consumer analysis and neuroscience applications [11], [12].

Near-range eye tracking is commonly used in wearable glasses devices, where near-infrared (NIR) LEDs are used to analyze the corneal reflections from the pupils and extract users' gaze information [13], [14]. In contrast, remote eye tracking is mainly required in applications such as AR 3D HUDs [4], [5], autostereoscopic 3D displays [8], [9], and DMS [3]. In these applications, remote eye tracking is important to rapidly and precisely detect and track the position of the user's eyes at distances of approximately 1 meter, even under varying lighting conditions [8]. Moreover, these eye tracking systems must be able to handle real-time processing and consider the constraints posed by limited system resources in vehicular embedded systems [4], [5]. Traditional remote eye tracking systems mainly rely on frame-based RGB cameras and NIR cameras, utilizing computer vision algorithms to identify and track the center position of the pupil [4], [5], [8], [9]. However, these methods often encounter challenges such as motion blur, limited frame rate speed, and sensitivity to lighting conditions [4], [5], [8], [9]. Fast eye movement detection is particularly important in eye tracking research to enable meaningful applications.

In recent years, asynchronous event image sensors [15], also known as dynamic vision sensors or neuromorphic cameras, have emerged as a promising alternative to traditional frame-based RGB cameras. Event cameras capture visual information in a fundamentally different way, detecting brightness changes asynchronously at the pixel level, resulting in an event stream that precisely captures moments when brightness variations occur [15], [16], [17]. Event cameras offer several advantages, including high dynamic range, high speed (up to around 10, 000fps), low power consumption, and efficient motion tracking, as they generate events only when actual changes occur [15], [16], [17]. Moreover, they are less sensitive to lighting variations and offer superior performance compared to frame-based RGB cameras, as shown in Figure 1. The unique characteristics of event cameras make them suitable for a wide range of emerging computer vision and robotics applications. Near-range eye tracking has been a popular use case for event cameras, particularly in wearable devices [18]. Some studies have performed remote eye tracking using event cameras, demonstrating the feasibility of face and eye detection using pretrained models on RGB face databases (DB) [19]. However, this study [19] did not conduct training using event camera data due to the challenges of constructing a large-scale face event camera image database.

In this paper, we aim to improve the performance of remote pupil tracking based on event cameras by constructing a large-scale event face training database using RGB-to-Event domain translation. Our contributions are as follows:

- Event Camera Training Data Generation: to generate event-like images from publicly available RGB face images, we adopted an image domain translation method, StyleFlow [20]. Our approach, which represents the first successful application of style-based image-to-image translation in this domain, enables the use of large-scale RGB datasets with existing annotations, such as key points and bounding boxes, to create event-style images without the need for manual labeling. Furthermore, we introduced a pre-processing technique that emphasizes the face region, reducing distortion caused by background information during the style translation process and resulting in more realistic event-like images. This method overcomes limitations of existing event image generation techniques that rely on video input citeb21 or image shaking [22], allowing for the generation of event-like images from a single static RGB image. This is particularly advantageous in scenarios where constructing video databases from face databases is difficult and time-consuming due to the extensive labeling process involved.

• Enhanced Pupil Localization on Real Event Camera Images: in our study, we conducted an evaluation of our proposed cross-modal learning-based pupil localization algorithm using real event camera images captured by DAVIS 346 (Inivation) [23]. These images were carefully selected to capture instances where the user's head movement was fast enough to reveal eye and face shape information. The training for this algorithm was conducted using a mixed dataset, incorporating both RGB and newly transformed event-like images, to implement a cross-modal learning approach. For pupil detection and localization, we utilized the RetinaFace algorithm [24], trained under this cross-modal learning strategy. Our approach yielded promising results, achieving a pupil localization accuracy of 97.2% on real event camera images, demonstrating its superiority over methods trained solely on RGB or event-like databases with transfer learning or learning from scratch. This successful outcome demonstrates the potential and effectiveness of remote event camera-based pupil localization in practical applications.

II. RELATED WORKS

A. EVENT CAMERA: BACKGROUND AND APPLICATIONS

The event camera is a next-generation image sensor that mimics the human visual system and possesses the following characteristics. Unlike frame-based cameras, it generates an asynchronous event stream, measuring pixel changes only when events occur, which are local changes in light intensity [15], [16], [17]. This feature provides high temporal resolution, resulting in a low camera system latency advantage and reduced motion blur for rapidly changing visual information [15], [16], [17]. When there is no change in light intensity, there is no event output, enabling it to operate with low power consumption. Another advantage is its ability to offer a high dynamic range (140dB), enabling it to adequately capture details in both dark and bright areas [23]. This proves especially beneficial in environments where light conditions change rapidly, such as driving through a tunnel.

Although event cameras are not yet commonly used in commercial applications, they are extensively applied and researched in various fields such as computer vision, image processing, AR/VR, robotics, and more. They are actively researched for object detection and tracking [25], pattern recognition [26], simultaneous localization and mapping (SLAM), and visual odometry [27] for real-time operation in limited embedded system resources such as vehicular and drones and more. In surveillance [28] and environmental monitoring [29] applications, their high dynamic range characteristic allows them to reliably monitor varying environments in low-light conditions and various lighting scenarios. Furthermore, event cameras can be used for depth [30] and optical flow estimation [31] by making use of multi-view event data to estimate 3D depth information and structure. Regarding the human-related computer vision field, they are researched in human pose estimation [32], hand tracking [33], near-range gaze tracking [18] for user interaction, human behavior understanding, and AR/VR wearable devices.

B. PUPIL TRACKING: RGB VS EVENT CAMERA-BASED APPROACHES

Previous studies on remote pupil tracking have mainly relied on frame-based RGB or NIR cameras [4], [5], [8], [9]. Frame camera-based eye pupil tracking algorithms often adopt machine learning techniques, where they first perform face area detection, then regress facial landmark points within the face region, and finally refine the pupil center position to track the center of the pupil [4], [5], [8], [9]. These studies utilize large-scale public face databases or face databases captured under various user and environmental conditions to train deep neural networks and complete their algorithms. It's important to note that there are several publicly available face datasets, such as CelebA [34], which includes 202, 599 face images from 10, 177 individuals, WIDER FACE [35], which has 32, 203 images, 300W [36], which contains 300 indoor and outdoor faces, as well as WFLW [37], with 10, 000 faces. These datasets have 5 to 98 facial landmark points and face bounding box annotations, which makes them valuable resources for facial detection and recognition research. These datasets are also actively used in remote pupil tracking studies.

The event camera offers specific advantages, which particularly include its speed and high dynamic range [15], [16], [17], that are well-suited for pupil localization. By using event cameras, fast eye movements that traditional frame-based pupil localization methods struggle to handle can be addressed, making it suitable for our primary applications in vehicular systems such as DMS and AR 3D HUD under various lighting conditions. Moreover, its low power consumption and low bandwidth consumption characteristics [15], [16], [17] align well with the limited system resources in vehicular systems. However, a significant challenge in utilizing event cameras for pupil localization research is the lack of publicly available face images and annotation datasets. Constructing and labeling diverse face databases with event cameras is a challenging task. While some datasets with limited numbers of face images captured by event cameras exist [38], they may have constraints in capturing clear shape information of the pupils due to the limited resolution of event cameras. Additionally, manual labeling is a precise and time-consuming process when real data needs to be captured [38]. To address this issue, various efforts have been made. One of the previous studies aimed to acquire event camera data for deep neural network-based algorithms using image shaking to create dynamic event streams [22]. However, there is a limitation in which facial annotations cannot be directly used. Other research focuses on generating event-like videos from video data [21].

However, the availability of face video data with keypoint annotations is limited.

In this paper, we present a solution to address this challenge. We conducted research on a method to generate event-like images from a publicly available RGB face database that includes all the necessary annotation information. To validate the effectiveness of the generated event-like training dataset, we trained a pupil localization algorithm using this dataset, utilizing our unique cross-modal learning strategy on a mixed dataset of both original RGB and the newly transformed event-like images and tested them on real event camera videos captured by DAVIS 346 [23]. The results demonstrated the effectiveness of our approach. One of the key advantages of our method is that it can generate event-like images from a single RGB image, making the labeling process more efficient compared to directly capturing or building video datasets, which can be challenging.

III. METHOD

We present a novel method for accurately locating the center of the pupil using an event camera. Specifically, we capture event images focusing on face images where the user's head movement is fast enough to reveal eye and face shape information. Our proposed method can be divided into two key steps: First, we generate training data for event-like images by an RGB-to-event image domain translation technique. For this, we utilize the StyleFlow [20] approach. To ensure the event-like images closely resemble real event camera images, we initiate a pre-processing stage on the content images used in this study. Second, we train the pupil localization model using the generated event-like images and the RetinaFace [24] algorithm, a state-of-the-art method for joint face detection and keypoint alignment. We then evaluate the performance of the trained model using real event images. The system overview is depicted in Figure 2. Our approach further improves pupil detection and alignment accuracy by creating a mixed training dataset that combines RGB images with the generated event-like images. Informed consent was obtained from all human subjects participating in this study.

A. EVENT-LIKE TRAINING IMAGE DATABASE GENERATION THROUGH RGB-TO-EVENT IMAGE DOMAIN TRANSLATION

Achieving high accuracy in pupil localization based on deep neural network algorithms necessitates a large-scale image database. For this purpose, we propose a methodology to create an event-like training image database, to enhance the accuracy of pupil localization in the context of event cameras. This method incorporates the StyleFlow [20] algorithm to convert RGB images into event-like images. StyleFlow is an algorithm that receives a content image and a style image as inputs and yields a target image that has been converted into the texture format of the style image while preserving the semantic information of the content image [20].

The process begins by extracting content features from the content image during a forward pass. After this, a backward pass is carried out, where the Style-Aware Normalization (SAN) module is used to perform a content-fixed style transformation. This transformation integrates the style features extracted from the style image, merging the characteristics of both content and style images [20]. Among the two inputs of StyleFlow, we use multiple pre-processed grayscale images obtained from RGB face images as content images. For the style image, a single real event camera image, captured directly by us, is used for learning. This process translates the RGB content image into an event-like image (target image) that closely resembles the real event camera image style image. Figure 3. (a) provides a visual representation of our proposed RGB-to-Event image domain translation method using our adaptive StyleFlow.

Figure 3. (b) shows multiple pre-processed content images and the results of event-like image generation for each pre-processing method. The first method converts the image to grayscale, as the event camera is a grayscale image that lacks color information. The second pre-processing method removes the background information and retains only the face regions to minimize the influence of unnecessary information on style translation. Based on the characteristics of event cameras, which measure only the amount of change in moving objects and maintain a constant gray intensity without capturing non-moving background information, the third method fills the background information in grayscale to mimic the event camera background.

We trained StyleFlow using these pre-processing methods. The content image database consisted of 70, 000 images from the public RGB CelebA [34] dataset, serving as the training content image database. A single real event face image, captured directly by us, was utilized as the style image database.

B. STRATEGIC TRAINING OF PUPIL LOCALIZATION MODELS USING EVENT CAMERAS: A CROSS-MODAL TRAINING APPROACH

We utilized the RetinaFace algorithm [24], a leading algorithm among multi-task joint learning strategies, for precise face detection and facial keypoint alignment in a single stage. RetinaFace is designed to predict the face bounding box and five face landmarks, including pupil centers, nose, and mouth points through training. It utilizes various backbones networks like ResNet, Mobilenet, and a pretrained network from ImageNet-11k [39]. As a lightweight model, RetinaFace can process VGA resolution images in real time on a single CPU core, making it suitable for real-time pupil tracking [24]. The aim of our research is to detect the exact location of the pupil center using the RetinaFace algorithm. We achieved this goal using a cross-modal training approach, integrating both RGB images and event-like images.

RGB images and event-like images each have their advantages, making their combination valuable in providing complementary information about face detection and pupil detection. This integration enhances the performance of the pupil localization model. RGB images effectively present the

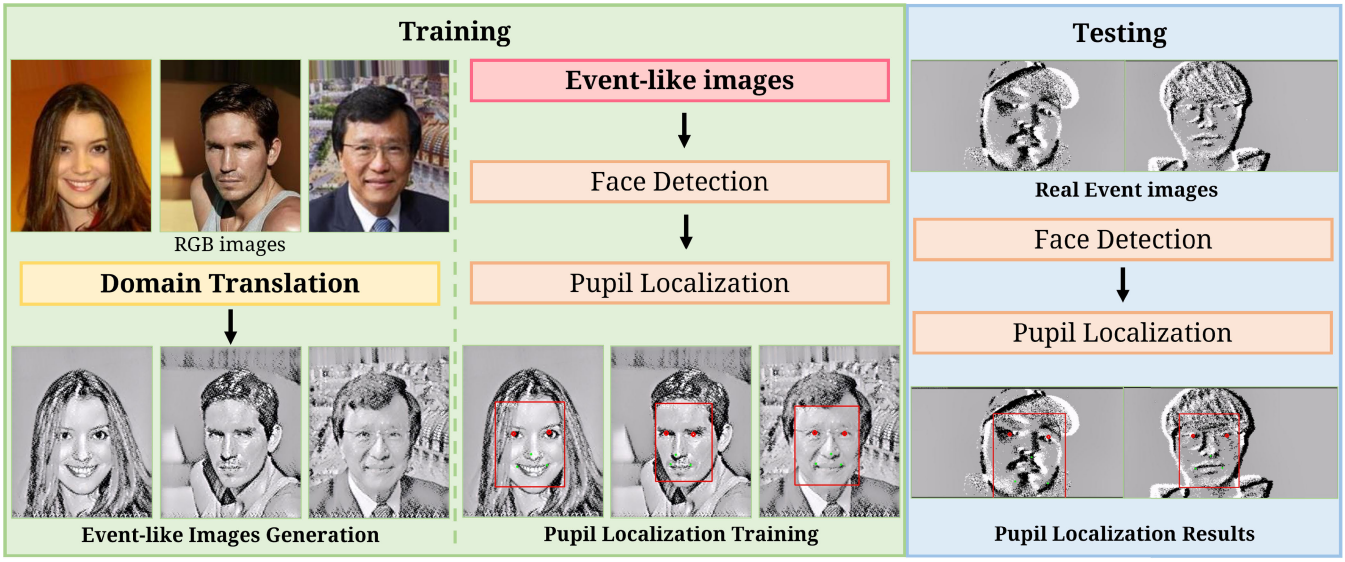

FIGURE 2. Overview of our proposed event camera-based pupil tracking method. During the training phase, we construct an event image training database by applying domain translation algorithms to convert RGB images into event-like images. Using the generated event-like images we train face detection and pupil localization algorithms. In the testing phase, we evaluate the performance of the trained algorithms on the real event camera images.

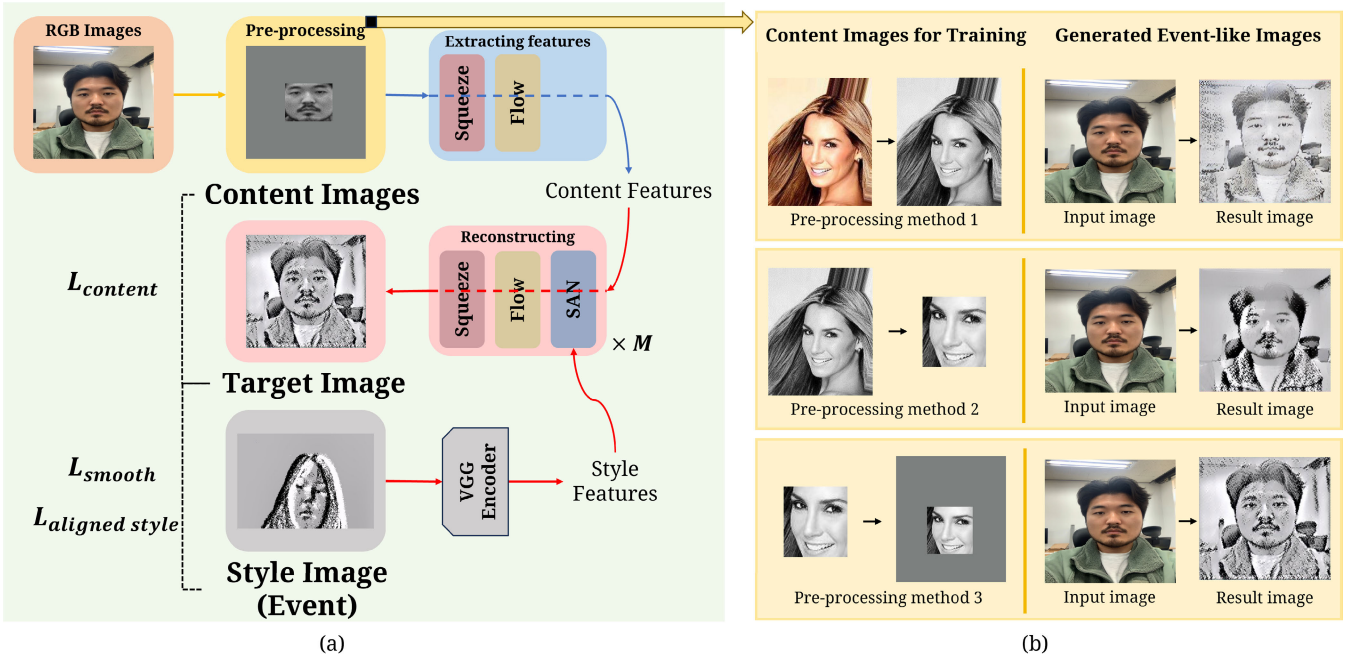

FIGURE 3. Illustration of our RGB-to-Event image domain translation using our adaptive StyleFlow [20] algorithm. (a) Multiple pre-processed frames serve as content image inputs with the background removed, retaining only face regions. These frames are also grayscale-converted. A singular real event camera image functions as the style input. Target images are then reconstructed considering content and style losses. (b) Comparative results of event-like image generation through StyleFlow [20], illustrating distinct content image pre-processing techniques: grayscale conversion, facial region cropping, and background fill with grayscale to emulate event camera background.

location of the pupil due to their sharp edge information, while event-like images capture pixel changes in a highly temporal and asynchronous manner, reflecting the rapid motion of the eyes, the fastest-moving organ in human faces [40]. However, when trained on event-like images only, the model could effectively identify the face bounding box but struggled with pupil alignment. This is because event-like images lack pupil shape information compared to RGB images, leading to imprecise keypoint alignment during pupil regression training. Additionally, real event camera data may contain variations and complexities not perfectly reflected in the event-like images generated using StyleFlow. Thus, cross-

modal training enables the model to adapt to a wider range of scenarios and perform well on real event camera data. By incorporating both modalities during training, the model effectively utilizes complementary information, leading to a more comprehensive understanding of the data.

TABLE 1. Training database for RGB-to-event image domain translation.

Training DB	DB Type	DB Number
Content DB	RGB CelebA [34]	202,599
Style DB	Real Event Face Image	1

For evaluation metrics, we used face detection accuracy and pupil alignment accuracy. Pupil alignment accuracy was measured by determining whether the difference between the predicted pupil localization and the ground truth was less than 10mm. We regarded such cases as successful. This was based on the assumption that the average human adult eye size is 24mm and the interpupillary distance (IPD) is 65mm, which allowed us to convert pixel errors to physical errors for measurement.

IV. EXPERIMENTAL RESULTS

In this research, we conducted a thorough evaluation of the two methods we proposed: 1) the RGB-to-Event image domain translation technique and 2) pupil localization using an event camera. Both algorithms were implemented in Python and trained and tested on an Ubuntu 20.04.6 LTS PC with an NVIDIA RTX 3090 (24GB) GPU. The final pupil localization algorithm was tested on real event camera images taken with a DAVIS 346 [23]. The metrics for evaluation, which were determined in the method section, were face detection accuracy and pupil alignment accuracy. These experiments allowed us to validate the practical performance of our proposed algorithms, and by comparing them with existing methods, assess the effectiveness of our proposed techniques. In this Experimental Results section, we will describe details of the datasets used and the performance of each proposed method.

A. EVALUATION - RGB-TO-EVENT IMAGE DOMAIN TRANSLATION

In this section, we evaluate the performance of our proposed RGB-to-Event Image Domain Translation technique. We utilized StyleFlow [20], which incorporated our proposed multiple pre-processing techniques, for generating Event-like images. The training dataset was constructed from 202, 599 RGB images from the CelebA [34] dataset, serving as content images through the proposed multiple pre-processing modules, and one real event camera image captured by DAVIS 346 [23], serving as the style image (Table 1). We selected one style image from multiple real event images that showed a fast-moving face, where the pupil information was clearly visible. This style image was selected through a trial-and-error process during training, as it yielded the best results. For training, we utilized the StyleFlow Pytorch implementation [41]. Our training employed the Ada optimizer, a learning rate of 0.00005, a batch size of 1, and a maximum of 70, 000 iterations. Other hyperparameters adhered to the default values stipulated in the implementation [41]. The process, which ran on our NVIDIA RTX 3090 (24GB) GPU, was completed in approximately three days.

The performance of the trained Event-image generation model was evaluated via a subjective visual test, which involved comparing the output images with real event camera images. However, due to the inherent limitations of subjective visual testing, the performance will be further evaluated based on the results of pupil localization in the subsequent subsection, focusing on results from the pupil localization task. The model is able to convert an RGB image with a resolution of 178 by 218 to an event-like image in approximately 2 seconds. As a result, we successfully constructed a database of 70, 000 event-like images, derived from RGB CelebA images, for use in training the pupil localization algorithm. Some examples of the event-like images generated by our model are depicted in Figure 4.

TABLE 2. Pupil localization training & testing database.

Training DB	DB Type	DB Number
RGB DB	WIDER FACE [35]	12.000
Event-like DB	Proposed RGB-To-Event Translation	7,500
Testing DB	**DB Type**	**DB Number**
Real Event DB	Captured With DAVIS 346 [23]	1,172

TABLE 3. Performance of the proposed pupil localization on real event images captured by DAVIS 346 [23]: face region detection accuracy and pupil alignment accuracy, considering a precision within 10mm.

Testing Images (Face Detection)	Success Number	Accuracy
Real Event Images	1165/1172	99.4%
Testing Images (Pupil Align)	**Success Number**	**Accuracy**
Bbox Detected Real Event Images	1132/1165	97.2%

B. EVALUATION - CROSS-MODAL TRAINING FOR EVENT CAMERA-BASED PUPIL LOCALIZATION

In this subsection, we evaluate the accuracy of our proposed model, trained with our cross-modal training method, in identifying the location of the pupil using an event camera. Our

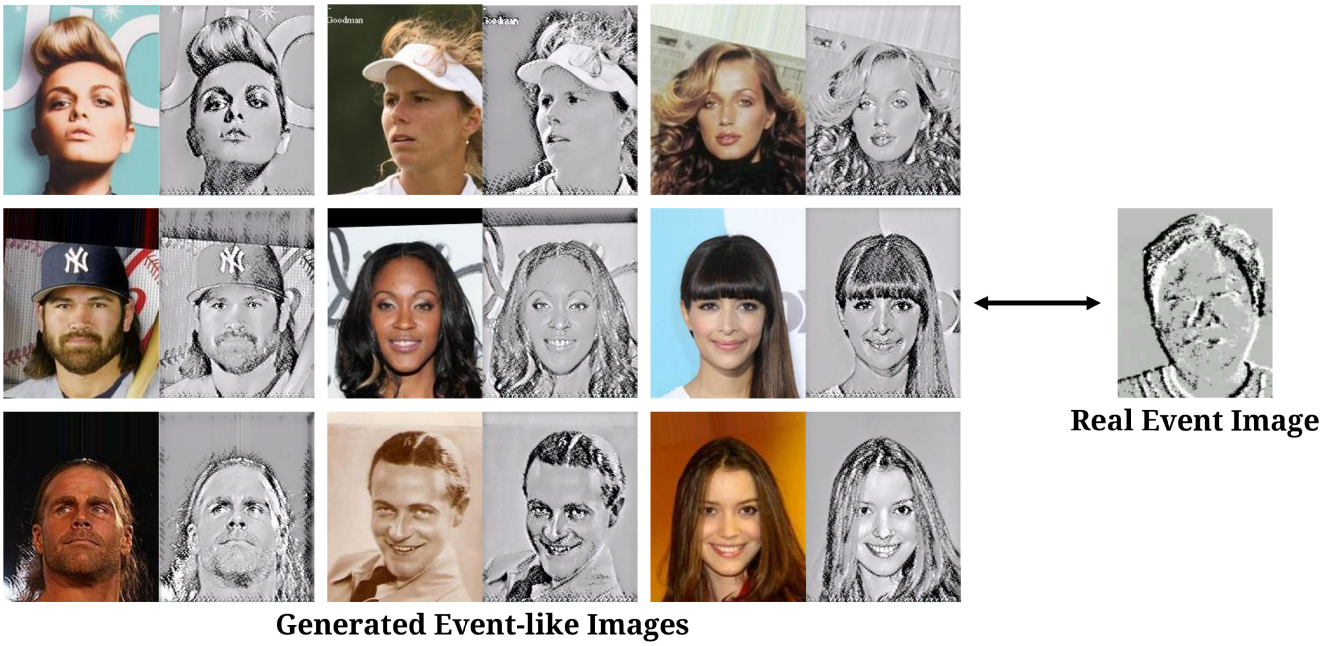

FIGURE 4. Experimental results of the proposed RGB-to-Event domain translation method. Odd columns display original RGB images, and even columns show the generated Event-like images. Real Event images captured by DAVIS 346 [23] are included for comparison purposes.

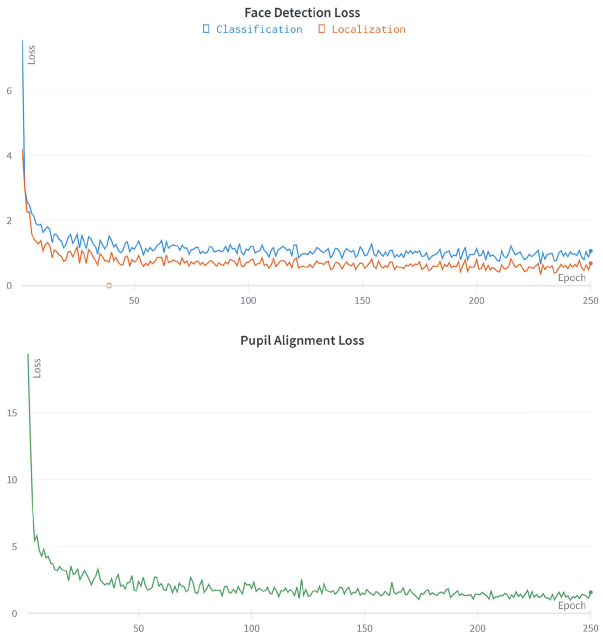

FIGURE 5. Learning curve of the proposed Event camera-based pupil localization: classification Loss, face bounding box loss, and pupil center localization loss across epochs. The x-axis represents the number of epochs, and the y-axis represents the corresponding losses.

FIGURE 6. Pupil localization results on event-Like images, generated from the proposed RGB-to-Event domain translation method. The red dots represent the pupil center positions, and the red boxes indicate the detected face regions.

combined training dataset comprises both RGB and event-like images, as indicated in Table 2. We used 12, 000 images with varying face sizes from the WIDER FACE dataset [35] for the RGB images. The event-like images consist of the 7, 500 images generated by our proposed RGB-to-Event image domain translation method, using CelebA [34] images for the RGB inputs. Both RGB and generated event-like images were resized to a resolution of 640 by 640 for training purposes. The RetinaFace [24] cross-modal training was conducted from scratch using PyTorch as implemented by [42]. The backbone network was MobileNet 0.25 [43], and we used a batch size of 128 for our training on two NVIDIA RTX 3090 (24GB) GPUs. The SGD optimizer was utilized for training, with other hyperparameters being set in accordance with the general training setup for RetinaFace [42]. The

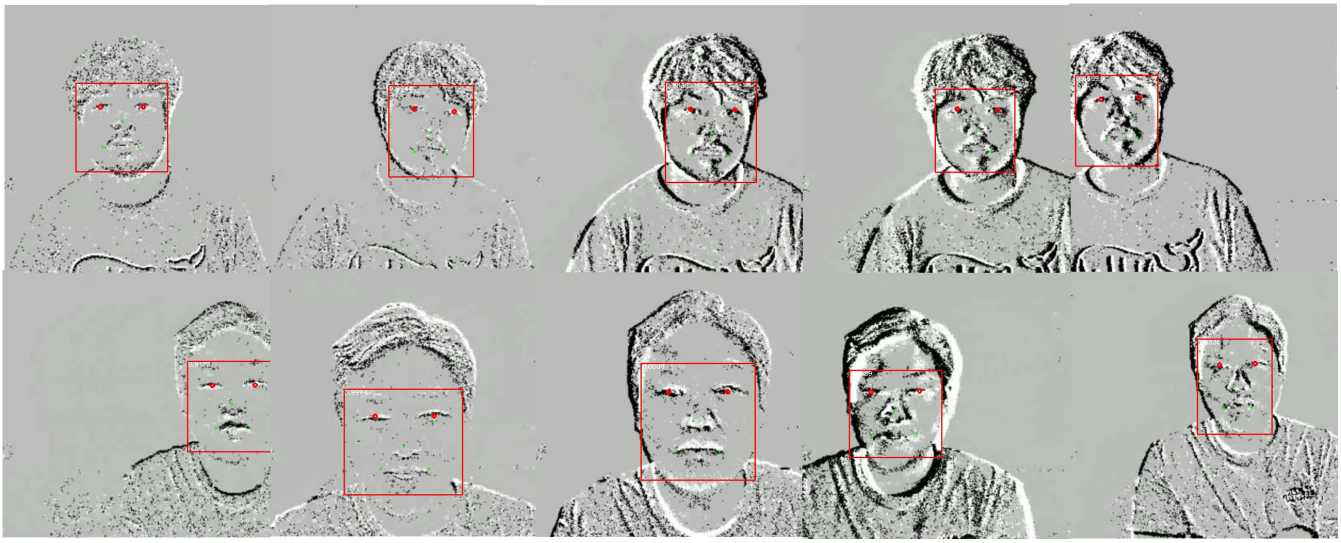

FIGURE 7. Proposed event camera-based pupil localization algorithm on the real event camera (DAVIS 346 [23]). Selected test images are displayed, capturing instances where the user's head movement is fast enough to reveal eye and face shape information using the DAVIS 346 [23] event camera. The red dots indicate the pupil center positions and the red boxes represent the detected face regions.

training process proceeded until the 250th epoch (Figure 5). The results of the pupil localization training are presented in Figure 6.

For testing, we utilized real event camera images captured by DAVIS 346 [23] with our cross-modal trained model. Two participants were recorded using DAVIS 346, with the participants instructed to move at a speed that would allow for clear capturing of face and pupil shape information. Given that our study aims to only test images that exhibit sufficient face and pupil information, we selected 1, 172 images with an adequate amount of edge detail from the video dataset for our testing (Table 2). The testing performance of the proposed pupil localization on the 1, 172 real event camera images demonstrated an accuracy of 99.4%, with successful face detection in 1, 165 instances. Among these successful face detection instances, pupil localization was successful in 1, 132 cases, indicating an accuracy of 97.2%. Importantly, we measured the precision of pupil localization against a threshold of 10 mm, based on the rationale provided in the Method section. This metric was chosen considering the average human adult eye size and interpupillary distance, converting pixel errors to physical errors for meaningful assessment (Table 3). Figure 7 provides a sample of the results obtained from the actual event camera, illustrating the successful localization of pupils within the 10 mm precision range.

V. DISCUSSION

Our results demonstrate the feasibility of pupil tracking based on event cameras. While the quality of the event-like images generated using our proposed method was acceptable according to subjective visual tests, their validity was also proved through the learning process for the pupil localization algorithm. Our final pupil localization algorithm showed strong performance when used with real event cameras where the user's head movement was fast enough to reveal eye and face shape information, indicating that our method for generating images similar to event images works effectively. The face detection accuracy was very high, at 99.4%, and the pupil alignment accuracy was 97.2%, a bit lower than detection but still showing excellent performance. As shown in the examples in Figure 7, even with various user movements, the system was able to accurately find the center of the pupil in the event camera images. This shows it could be suitable for vehicular systems like DMS and AR 3D HUD applications where the driver's face may move in various ways.

In the RGB to Event image domain translation, we successfully produced event-like images very similar to actual event camera images using our adaptive StyleFlow algorithm, which combines multiple pre-processing methods and the StyleFlow [20] algorithm. This achievement represents the first successful application of style-based image-to-image translation specifically for event camera image generation, marking a significant advancement in this field. To further evaluate the superiority of our method, we extended our comparisons to include not only the Style Transfer [44] algorithm, one of the most renowned methods in style translation, but also additional real-time style transfer approaches like Fast Style Transfer [45] and AdaIN [46]. The Fast Style Transfer algorithm, based on a perceptual loss [47], and AdaIN, which transfers the global mean and variance of a style image to a content image in the feature space, were both trained and tested with an event image style. These additional methods provide a comprehensive perspective on the adaptability and efficiency of various style translation techniques in event-like image generation. We also conducted training and performance comparison tests with the original

FIGURE 8. Comparative visualization of RGB-to-Event image domain translation methods. From left to right: Original RGB images (first column), images translated using Style Transfer [44] (second column), fast style transfer [45] (third column), AdaIN [46] (fourth column), StyleFlow [20] (fifth column), and our adaptive StyleFlow enhanced with the proposed pre-processing method (sixth column). The seventh column displays real event images captured by DAVIS 346 [23] for direct comparison.

TABLE 4. Performance of the proposed pupil localization method according to different training strategies.

Training Strategy (Face Detection)	Training DB	DB Number	Test Accuracy (Face Detection)
Pretrained	Pretrained (RGB)	N/A	49.7% (583/1,172)
Transfer Learning	Pretrained (RGB) + Event-like DB	1,000	93.9% (1,100/1,172)
Transfer Learning	Pretrained (RGB) + Event-like DB	7,500	98.9%(1159/1172)
Learning From Scratch	Event-like DB	32,000	96.8% (1,159/1,172)
Learning From Scratch	Event-like DB	70,000	81.9% (960/1,172)
Cross-modal Learning	RGB + Event-like DB	12,000 + 7,500	**99.4% (1,165/1,172)**
Training Strategy (Pupil Align)	**Training DB**	**DB Number**	**Test Accuracy (Pupil Align)**
Pretrained	Pretrained (RGB)	N/A	21.8% (127/583)
Transfer Learning	Pretrained (RGB) + Event-like DB	1,000	38.0% (418/1,100)
Transfer Learning	Pretrained (RGB) + Event-like DB	7,500	11.2% (130/1,159)
Learning From Scratch	Event-like DB	32,000	55.7% (632/1,135)
Learning From Scratch	Event-like DB	70,000	4.69% (45/960)
Cross-modal Learning	RGB + Event-like DB	12,000 +7,500	**97.2% (1,132/1,165)**

StyleFlow [20] algorithm to demonstrate the superiority of our proposed pre-processing module. All of these previous methods, including the original StyleFlow [20], used real event camera images as style input. However, when generating event-like images, the Style Transfer [44], Fast Style Transfer [45], and AdaIN [46] algorithms resulted in images that, upon pixel-by-pixel comparison with real event camera images, appeared more clustered and less reflective

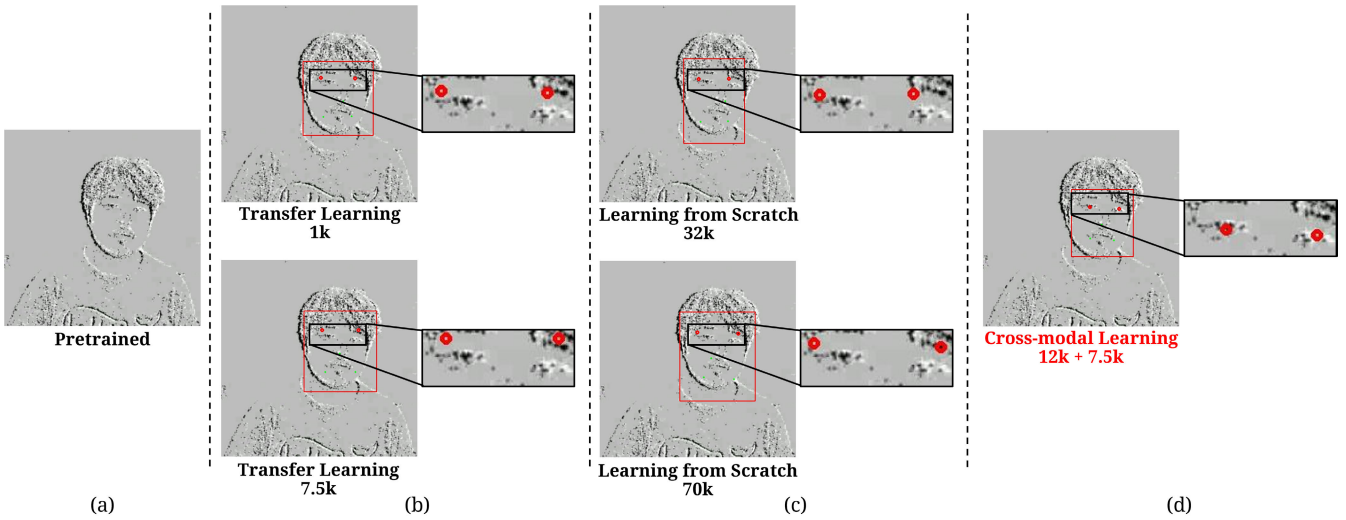

FIGURE 9. Comparison of pupil localization results using different strategies, evaluated with DAVIS 346 images from [23]. The red box indicates the detected face box, and the two red circles show the localized pupil centers in each figure. Illustrated models are: (a) pretrained on RGB WiderFace [35], (b) transfer learning with 1k and 7.5k event-like sets, (c) learning from scratch on 32k and 70k event-like sets, and (d) our "Cross-modal Learning" approach, which utilized learning from scratch by integrating 12k RGB and 7.5k event-like images."

of the distinct pixel characteristics inherent in event camera images. While the original StyleFlow [20] algorithm reduced this clustering issue somewhat, it was not removed entirely, and a variety of artifacts appeared in the facial images due to background information. When we used our proposed multiple pre-processing method and StyleFlow [20], these artifacts disappeared because the focus was on converting the face region to an event camera, resulting in images very similar to real event images. As a result, our proposed method was successful in generating event-like images without artifacts, capturing the sensitive, sparse responses to pixel-level motion, a characteristic of event cameras. A comparison of event-like image generation results for each method is shown in Figure 8.

To validate the effectiveness of our proposed cross-modal training, we compared the results of transfer learning on a pretrained RetinaFace model [42] using our event-like database, learning from scratch using only event-like data, and our proposed cross-modal training. All tests were performed on the same 1, 172 real event images. Table 4 shows the performance comparison of these different training methodologies. In terms of face detection accuracy, both transfer learning and learning from scratch showed a significant improvement over the pretrained model based on RGB images. However, when it came to pupil alignment, neither transfer learning nor learning from scratch, despite using the event-like image database, demonstrated any significant performance boost. Only our proposed cross-modal learning methodology achieved a high accuracy of 97.2%. Figure 9 describes the visual comparison of pupil localization results by different training strategies. We measured success accuracy based on a precision threshold of 10 mm for pupil localization. Other training strategies often resulted in an output with errors greater than 10 mm from the center of the pupil and exceeded the boundaries of the eye shape. This emphasizes the effectiveness of our cross-modal training methodology, which uses a mixed database of RGB and event-like images. One of the noticeable aspects of Figure 9 is how the tightness of the face detection bounding box affects the precision of pupil localization. The bounding box results created by our proposed cross-modal learning were tighter compared to other methods, leading to an increase in pupil localization precision. This indicates that while other methods could detect faces with high success rates, their box regression wasn't accurate, which in turn affected the final pupil localization precision. This suggests that our cross-modal training method, utilizing a mixed database of RGB and event-like images, can adapt to a wider range of scenarios and perform well on real event camera data. By incorporating both modalities during training, the model effectively utilizes complementary information, leading to a more comprehensive understanding of the data. Despite the positive results, our study also has limitations. The performance was lowest when learning solely from event-like images, indicating imperfections in our RGB-to-event image domain translation. This observation is strengthened as we expanded the dataset for such learning from 32, 000 to 70, 000 images. This increase in the size of the training data led to a decrease in both detection accuracy and pupil alignment accuracy.

A. COMPARISON WITH EXISTING APPROACHES

Our proposed study is aligned with existing research efforts in the field of eye tracking using event cameras. Event cameras have been extensively studied for near-range eye-gaze tracking, primarily for wearable devices, where high temporal resolution is essential. For instance, Angelopoulos [18] combined NIR cameras with event cameras to

TABLE 5. Performance comparison with other methods.

Model	Near / Remote	Test DB (Event Sensor)	Test DB (Number)	Training DB (Source)	Training DB (Number)	Face Detection	Eye-region Detection	Pupil Center Localization	Additional Performance
[19]	Remote	Prophesee Gen3 ATIS [48]	10 Subject Video (raw event)	N/A	N/A	59% (Accuracy)	59% (Accuracy)	N/A	Blink Detection : 59% (Accuracy)
[49]	Remote	Prophesee Gen4 [50]	3 Subject Video (event frame)	Synthetic (event frame)	2000 Video	90% (Precision)	90% (Precision)	N/A	Blink Detection : 88.93% (Precision)
[53]	Remote	DAVIS 346 [23]	26 Subject Video (event frame)	Real Event (event frame)	130 Videos	N/A	N/A	N/A	Drowsiness Detection : 97.2% (Accuracy)
[38]	Remote	Prophesee Gen4 [50]	6 Subject Video (event frame)	Synthetic (event frame)	600 Videos	N/A	N/A	N/A	Facial Expression : 30.95% (Accuracy)
Ours	Remote	DAVIS 346 [23]	2 Subject Video (event frame)	Synthetic (event frame) + RGB	12,000 images + 7500 images	**99.4% (Accuracy) 99% (Precision)**	N/A	**97.2%**	N/A

achieve higher temporal resolution in gaze tracking compared to traditional NIR camera-based methods. This approach used event cameras for temporal interpolation between NIR image frames, providing a solution to capture fast eye movements that conventional RGB or NIR cameras may miss.

However, when it comes to remote eye tracking using event cameras, research is still in its early stages. Lenz et al. [19] focused on remote eye blink detection using Prophesee Gen3 ATIS [48]. This method used the temporal signature of eye blinks to detect whether blinks occurred. Based on this information, a Gaussian tracker and face detector were utilized, enabling face detection only when blinks were detected. While this study provided insights into face detection and blink-related eye-region detection, it did not include pupil center localization, a core aspect of our research. Ryan et al. [49] also aimed at remote blinking detection using Prophesee Gen4 [50] and achieved high precision in face and eye-region detection. To enable face and eye-region detection, they proposed a network architecture called GR-YOLO. The training process utilized synthetic event frame images generated from N-Helen data. The N-Helen dataset was created by applying random augmentations and transformations with 6 degrees of freedom (DOF) to RGB Helen data [51], using a deep learning-based frame interpolation algorithm to generate video data. This generated video data was then transformed into synthetic event data using the ESIM [52] algorithm, a video-to-event simulator. Face and eye-region detection achieved a precision performance of 90%. However, they did not undertake pupil center localization as a separate task. Furthermore, EDDD method, as presented by [53], primarily focuses on event-based drowsiness detection for driving safety. They mounted a DAVIS 346 [23] event camera in vehicles to collect a significant dataset of event camera images. Using this dataset, they trained machine learning models for drowsiness detection with high accuracy, ranging from 94.42% to 99.9%, across various scenarios. Notably, this paper did not evaluate face detection, eye detection, or pupil center localization accuracy. The NEFER method [38] focused on creating a facial expression dataset using event cameras, generating synthetic event data from GoPro-captured high-resolution RGB video data through ESIM [52]. They trained a YOLOv2 [54]-based event-camera face detector and developed Xception [55]-based face alignment techniques. While they achieved a facial expression accuracy of 30.95%, they did not assess the accuracy of real event camera-based face, eye, or pupil detection. Please refer to Table 5 for a detailed performance comparison.

One of the primary advancements in our work is the accurate localization of pupil centers. This is particularly significant as pupil center localization plays an essential role in AR 3D HUD applications [4]. Accurately localizing pupil centers is more intricate than detecting face or eye regions, especially considering the subtle shape variations in the entire face region. Additionally, our RGB-to-event image domain translation method, StyleFlow [20], enables efficient generation of synthetic event training data from a single RGB image, eliminating the need for extensive video datasets. Unlike methods relying on video input or image shaking to create dynamic event streams, our approach efficiently produces event-like images from static RGB images, reducing the complexity and time required for dataset construction. Furthermore, our cross-modal learning strategy, utilizing both RGB and event images, significantly enhances the performance of face detection and pupil localization compared to previous studies. We outperformed existing methods in terms of accuracy. In summary, our study demonstrates substantial advancements in remote pupil tracking using event cameras, with a focus on accurate pupil center localization. We have overcome the challenges of previous studies, presenting an efficient method for creating synthetic event training data and proposing cross-modal learning techniques.

VI. CONCLUSION

In conclusion, our study successfully demonstrates the potential of event camera-based pupil tracking. The method we proposed for generating event-like images delivered promising results with visuals closely matching real event camera images. Additionally, the pupil localization training with these generated event-like images proved the effect of the RGB-to-event image domain translation. Our final pupil localization algorithm, developed with our proposed cross-modal training strategy, demonstrated high accuracy in both face detection and pupil alignment. This surpassed the performance of previous deep learning-based methods trained on conventional RGB images and traditional training schemes. Our approach, representing the first successful application of style-based image-to-image translation for event camera image generation, establishes a new foundation in synthetic event image generation. Additionally, recognizing the potential for network architecture advancements specifically customized for event camera data, we have outlined this promising direction as a future research area, expanding on the foundation established by our current work. In sum, our work marks a significant step forward in event camera-based pupil tracking technology.

REFERENCES

[1] X. Zhang and S.-M. Yuan, "An eye tracking analysis for video advertising: Relationship between advertisement elements and effectiveness," *IEEE Access*, vol. 6, pp. 10699–10707, 2018.

[2] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 5048–5054.

[3] M.-C. Chuang, R. Bala, E. A. Bernal, P. Paul, and A. Burry, "Estimating gaze direction of vehicle drivers using a smartphone camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 165–170.

[4] 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, doi: 10.1109/ACCESS.2021.3110644.

[5] D. Kang and H. S. Chang, "Low-complexity pupil tracking for sunglasses-wearing faces for glasses-free 3D HUDs," *Appl. Sci.*, vol. 11, no. 10, p. 4366, May 2021, doi: 10.3390/app11104366.

[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, doi: 10.1145/3206343.3206349.

[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, Nov. 2016, Art. no. 179.

[8] D. Kang and J. Heo, "Content-aware eye tracking for autostereoscopic 3D display," *Sensors*, vol. 20, no. 17, p. 4787, Aug. 2020.

[9] D. Kang, J.-H. Choi, and H. Hwang, "Autostereoscopic 3D display system for 3D medical images," *Appl. Sci.*, vol. 12, no. 9, p. 4288, Apr. 2022, doi: 10.3390/app12094288.

[10] D. Rozado, T. Moreno, J. San Agustin, F. B. Rodriguez, and P. Varona, "Controlling a smartphone using gaze gestures as the input mechanism," *Hum.-Comput. Interact.*, vol. 30, no. 1, pp. 34–63, Jan. 2015, doi: 10.1080/07370024.2013.870385.

[11] S. Bottos and B. Balasingam, "Tracking the progression of reading using eye-gaze point measurements and hidden Markov models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7857–7868, Oct. 2020, doi: 10.1109/TIM.2020.2983525.

[12] L. Dai, J. Liu, and Z. Ju, "Binocular feature fusion and spatial attention mechanism based gaze tracking," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 2, pp. 302–311, Apr. 2022.

[13] C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," *Comput. Vis. Image Understand.*, vol. 98, no. 1, pp. 4–24, Apr. 2005.

[14] 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, doi: 10.1109/TBME.2007.895750.

[15] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[16] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008, doi: 10.1109/JSSC.2007.914337.

[17] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240 × 180 130 dB 3 μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Oct. 2014, doi: 10.1109/JSSC.2014.2342715.

[18] A. N. Angelopoulos, J. N. P. Martel, A. P. S. Kohli, J. Conradt, and G. Wetzstein, "Event based, near eye gaze tracking beyond 10, 000Hz," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 5, pp. 2577–2586, May 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9389490/

[19] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Frontiers Neurosci.*, vol. 14, p. 587, Jul. 2020, doi: 10.3389/fnins.2020.00587.

[20] W. Fan, J. Chen, J. Ma, J. Hou, and S. Yi, "StyleFlow for content-fixed image to image translation," 2022, *arXiv:2207.01909*.

[21] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.

[22] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3583–3592.

[23] G. Taverni, D. P. Moeys, C. Li, C. Cavaco, V. Motsnyi, D. S. S. Bello, and T. Delbruck, "Front and back illuminated dynamic and active pixel vision sensors comparison," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 5, pp. 677–681, May 2018, doi: 10.1109/TCSII.2018.2824899.

[24] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," 2019, *arXiv:1905.00641*.

[25] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–9.

[26] D. P. Moeys, F. Corradi, E. Kerr, P. Vance, G. Das, D. Neil, D. Kerr, and T. Delbrück, "Steering a predator robot using a mixed frame/event-driven convolutional neural network," in *Proc. 2nd Int. Conf. Event-Based Control, Commun., Signal Process. (EBCCSP)*, Jun. 2016, pp. 1–8.

[27] J. Hidalgo-Carrió, G. Gallego, and D. Scaramuzza, "Event-aided direct sparse odometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5771–5780.

[28] T. Bolten, R. Pohle-Fröhlich, and K. D. Tönnies, "DVS-OUTLAB: A neuromorphic event-based long time monitoring dataset for real-world outdoor scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1348–1357.

[29] G. Chen, P. Liu, Z. Liu, H. Tang, L. Hong, J. Dong, J. Conradt, and A. Knoll, "NeuroAED: Towards efficient abnormal event detection in visual surveillance with neuromorphic vision sensor," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 923–936, 2021, doi: 10.1109/TIFS.2020.3023791.

[30] M. Muglikar, L. Bauersfeld, D. P. Moeys, and D. Scaramuzza, "Event-based shape from polarization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1547–1556.

[31] L. Pan, M. Liu, and R. Hartley, "Single image optical flow estimation with an event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1669–1678.

[32] S. Zou, C. Guo, X. Zuo, S. Wang, P. Wang, X. Hu, S. Chen, M. Gong, and L. Cheng, "EventHPE: Event-based 3D human pose and shape estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10976–10985.

[33] R. Page, "Live demonstration: Integrating event based hand tracking into TouchFree interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4033–4034.

[34] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[35] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.
[36] 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*, Dec. 2013, pp. 397–403.
[37] 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.*, Jun. 2018, pp. 2129–2138.
[38] L. Berlincioni, L. Cultrera, C. Albisani, L. Cresti, A. Leonardo, S. Picchioni, F. Becattini, and A. Del Bimbo, "Neuromorphic event-based facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4108–4118.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[40] R. Vertegaal, "A Fitts law comparison of eye tracking and manual input in the selection of visual targets," in *Proc. 10th Int. Conf. Multimodal Interfaces*, Oct. 2008, pp. 241–248, doi: 10.1145/1452392.1452443.
[41] W. Fan. (Jul. 2022). *StyleFlow For Content-Fixed Image to Image Translation*. [Online]. Available: https://github.com/weepiess/StyleFlow-Content-Fixed-I2I
[42] RetinaFace in PyTorch Contributors. (Apr. 2020). *RetinaFace in PyTorch*. [Online]. Available: https://github.com/biubug6/Pytorch_Retinaface
[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4510–4520.
[44] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.
[45] Logan Engstrom. (Oct. 2016). *Fast Style Transfer*. [Online]. Available: https://github.com/lengstrom/fast-style-transfer
[46] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 1501–1510.
[47] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[48] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, Jan. 2011.
[49] C. Ryan, B. O'Sullivan, A. Elrasad, A. Cahill, J. Lemley, P. Kielty, C. Posch, and E. Perot, "Real-time face & eye tracking and blink detection using event cameras," *Neural Netw.*, vol. 141, pp. 87–97, Sep. 2021, doi: 10.1016/j.neunet.2021.03.019.
[50] T. Finateu, A. Niwa, D. Matolin, K. Tsuchimoto, A. Mascheroni, E. Reynaud, P. Mostafalu, F. Brady, L. Chotard, F. LeGoff, H. Takahashi, H. Wakabayashi, Y. Oike, and C. Posch, "A 1280 × 720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 μm pixels, 1.066 GEPS readout, programmable event-rate controller and compressive data-formatting pipeline," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 112–114.
[51] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "'Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2012, vol. 7574, no. 3, pp. 679–692.
[52] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. Conf. Robot Learn.*, 2018, pp. 969–982.
[53] 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, doi: 10.1109/JSEN.2020.2973049.
[54] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[55] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

DAEHYUN KANG received the B.S. degree from the School of Electronic and Electrical Engineering, Hongik University, Seoul, South Korea, in 2023, where he is currently pursuing the master's degree. His research interest includes neuromorphic image sensors.

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. From 2013 to 2021, he was a Senior Researcher with the Samsung Advanced Institute of Technology, Suwon, South Korea. In 2021, he joined the School of Electronic and Electrical Engineering, Hongik University, Seoul, where he is currently an Assistant Professor. His research interests include image processing and computer vision including detection, tracking, segmentation, image enhancement, application to augmented reality, autostereoscopic 3D displays, and medical image analysis.