

Real-Time Gaze Tracking with Event-Driven Eye Segmentation

Yu Feng*
University of Rochester

Nathan Goulding-Hotta†
Reality Labs Research

Asif Khan‡
Reality Labs Research

Hans Reyserhove§
Reality Labs Research

Yuhao Zhu¶
University of Rochester

ABSTRACT

Gaze tracking is increasingly becoming an essential component in Augmented and Virtual Reality. Modern gaze tracking algorithms are heavyweight; they operate at most 5 Hz on mobile processors despite that near-eye cameras comfortably operate at a real-time rate (> 30 Hz). This paper presents a real-time eye tracking algorithm that, on average, operates at 30 Hz on a mobile processor, achieves 0.1° – 0.5° gaze accuracies, all the while requiring only 30K parameters, one to two orders of magnitude smaller than state-of-the-art eye tracking algorithms. The crux of our algorithm is an Auto ROI mode, which continuously predicts the Regions of Interest (ROIs) of near-eye images and judiciously processes only the ROIs for gaze estimation. To that end, we introduce a novel, lightweight ROI prediction algorithm by emulating an event camera. We discuss how a software emulation of events enables accurate ROI prediction without requiring special hardware. The code of our paper is available at <https://github.com/horizon-research/edgaze>.

Keywords: Gaze, eye tracking, event camera, segmentation

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed/augmented reality; Computing methodologies—Computer graphics—Shape modeling—Shape analysis

1 INTRODUCTION

Gaze tracking is critical to many fields [51], such as medicine [9, 11], human-machine interaction [14, 39, 52, 61], psychology [53], augmented reality (AR) [38, 57], and virtual reality (VR) [18, 36, 56, 66, 69]. Due to its importance, there has been much research improving the tracking accuracy, both using conventional infrared/RGB cameras [15, 40, 43, 46, 68] and emerging event cameras [8].

The speed of gaze tracking is critical. Gaze tracking, in many cases, directly interfaces with users, and is often just the first stage of an end-to-end application. Studies have shown that a tracking speed of at least 30 Hz is required for real-time user-facing applications such as VR [7]. Today’s eye tracking systems, however, are generally an order of magnitude slower than the requirement. For instance, on Nvidia’s Jetson Xavier, a powerful mobile computing platform, two state-of-the-art algorithms, RITNet [15] and DeepVOG [68], operate at a mere 5 Hz (while achieving a sub- 0.5° gaze error).

This paper demonstrates a gaze tracking system that achieves 30 Hz tracking speed on a mobile GPU with a sub- 0.5° gaze accuracy. Most prior eye tracking algorithms work on a frame-by-frame basis. However, eye tracking in virtually all use-cases processes continuous

frames, which exposes temporal information that is often ignored. Our system leverages this temporal correlation for real-time tracking.

Event-Driven Auto ROI In particular, we use the temporal correlation across frames to predict the Region of Interest (ROI) in the current frame. We perform processing only in the ROIs whenever possible, falling back to full-resolution processing very rarely (0.6% of the time). The ROIs usually contain only 18%–32% of the pixels compared to full resolution, greatly improving the execution speed.

Our contribution here is a novel ROI prediction algorithm specialized for eye tracking. In particular, we emulate an event camera to extract the temporal correlation across frames, which then guides a lightweight deep neural network (DNN) to predict the ROI. The ROI prediction network essentially gives rise to an *Auto ROI mode* for eye tracking, akin to the conventional 3A algorithms for photography (auto focus, auto white balance, and auto exposure). Similar to the 3A algorithms, this “fourth A” mode operates in a feedback-driven fashion, where we use information extracted from the current frame (e.g., events, edges, ROI) to predict the ROI of future frames.

Why Events? Event cameras report only the changes in pixel brightness level, a.k.a., events. We use events for ROI prediction for two reasons. First, events inherently encode the temporal changes, and thus provide natural cues to ROI prediction. Second, events are naturally sparse and can be efficiently encoded, enabling a small neural network with a low overhead in ROI prediction. Computation for the ROI prediction must be lightweight, in order to not offset the gains from processing only ROI images.

Why Not an Event Camera? We do not directly use an event camera for two main reasons. First, event cameras are not readily available in many real-world systems (e.g., VR) and usually require dedicated downstream algorithms, which limit their applicability. Second, although event cameras can operate at a frequency as high as 10K Hz, useful for capturing precise eye movement, such a high speed is usually an overkill for many consumer eye-tracking applications, e.g., user interface navigation in VR, where the target speed is “only” 30 Hz, easily achievable with conventional cameras.

Efficiency-Oriented Segmentation Coupled with the Auto ROI mode, we propose a lightweight neural network-based algorithm to predict the gaze from an ROI. Like much of the recent work [15, 43, 46, 68], we use a model-based approach, where we first parameterize an eye image through eye segmentation, which is then used to estimate the gaze through fitting a geometric eye model.

We propose a new segmentation network designed with efficiency in mind. We show that depth-wise separable convolution with wider channels provides a segmentation backbone that balances accuracy and compute complexity. Our network requires 13 – $93\times$ fewer parameters (30K in total) and 20 – $31\times$ fewer arithmetic operations than existing DNN algorithms (RITnet [15] and DeepVOG [68]).

Our dataset and code will be released. Our contributions are:

- We introduce *event-driven Auto ROI* for eye tracking, a novel event-driven ROI prediction algorithm for eye tracking using software-emulated events and temporal feedback.
- We propose an accurate eye segmentation neural network co-trained with ROI prediction; the segmentation network is an order of magnitude simpler compared to prior algorithms.

*e-mail: yfeng28@ur.rochester.edu

†e-mail: nrgh@fb.com

‡e-mail: aikhan@fb.com

§Now at Impinj, e-mail: hans.reyserhove@impinj.com

¶e-mail: yzhu@rochester.edu

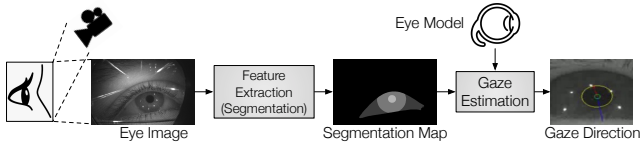


Fig. 1: A typical model-based eye tracking pipeline, which consists of a feature extraction stage and a gaze estimation stage. Feature extraction is usually done through eye segmentation. Gaze estimation is much more lightweight than feature extraction.

2 RELATED WORK

Eye Tracking Algorithms Eye tracking methods fall into two main categories: model-based methods and appearance-based methods [31, 70]. Appearance-based methods directly learn a mapping from an eye image to the gaze directions [50, 67, 71, 72]. We focus on the model-based method, which predicts gazes using a physiology-inspired eye model. Model-based methods are generally regarded as providing better accuracy than appearance-based methods [70].

Figure 1 illustrates the pipeline of a typical model-based method [48, 68], which consists of two stages: 1) eye feature extraction and 2) gaze estimation. Feature extraction algorithms take near-eye images as inputs and extract salient features of the eye. The features represent a parameterization of the eye. The parameters are then used to fit a geometric eye model to estimate the gaze. The eye model is a physics-based 3D model typically pre-constructed and calibrated from experimental data [16, 34, 48, 68, 74].

Compared to gaze estimation, feature extraction is much more time consuming. Our experiments show that feature extraction contributes to 85% of the total execution time across a range of different algorithms, and is thus a prime target for optimization. We introduce a lightweight neural network to extract features through segmentation, which reduces the computation complexity by an order of magnitude compared to state-of-the-art networks.

Feature Extraction While historically hand-crafted, geometric features are popular [1, 26, 32, 62], DNN algorithms have recently been shown to be more robust and accurate [15, 25, 43, 46, 48, 68]. In particular, extracting features through eye segmentation [15, 43, 46, 68] is by far the most widely used method.

An eye segmentation algorithm usually classifies the pixels in an eye image into different classes. The resulting segmentation map has the same dimensions as the eye image, and each pixel value in the map represents the class ID of that pixel. Our algorithm uses a similar approach as DeepVOG [68], RITnet [15], OpenEDS [5], and Kim et al. [43], known as “part segmentation,” which segments the eye image into four parts: pupil, iris, sclera, and background. In contrast, Ellseg [46] and Wang et al. [64] use elliptical segmentation, which predicts the ellipses of the pupil and iris.

Prior studies focus primarily on accuracy. In contrast, we focus on *compute efficiency*, and show that tracking can be five times faster with little sacrifice in accuracy. We note that while we demonstrate our ROI prediction on part segmentation-based eye tracking, the idea applies generally to tracking algorithms using other features.

Event Camera Event cameras operate each pixel independently and asynchronously [27]. Each pixel gets activated when its intensity change surpasses a predefined threshold. The response is called an “event”, which includes the pixel coordinates, a timestamp, and a polarity value. Because of the increased hardware complexity, the resolution of event cameras is typically lower than that of conventional cameras [27], but event cameras have a much higher frame rate (upward of tens of thousands of Hz) since they produce only (sparse) events occasionally rather than (dense) pixels regularly.

Event cameras appear in a range of vision and robotics tasks such as object tracking [58], localization [65], and reconstruction [42]. Recent work has also started using event cameras for eye track-

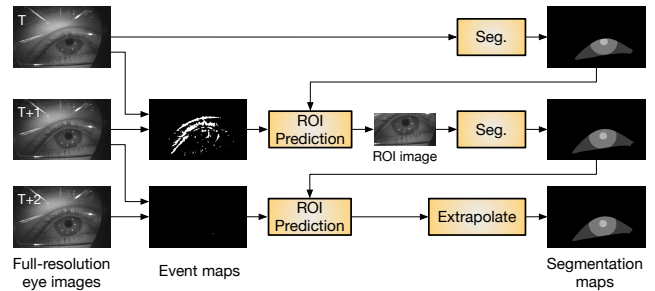


Fig. 2: Overview of our event-driven eye segmentation. Time progresses from top to bottom. The segmentation results are used by a common gaze estimation algorithm [63], which is omitted in the figure. We generate an event map from every two consecutive eye frames; the current event map and previous segmentation map are combined to predict the ROI. If the number of events is small (e.g., shown at time $(T + 2)$), we can simply extrapolate the segmentation result rather than performing a full-blown segmentation.

ing [8, 19], and is able to achieve a 10K Hz frequency. While extremely high tracking frequency is needed when capturing precise eye movement (e.g., foveated rendering), a lower frame rate provided by conventional cameras is sufficient for many eye tracking use cases, e.g., eye communication system for disability [12]. Instead of using event camera hardware, we emulate event camera output from our conventional sensor using software.

ROI Computing ROI is widely used to reduce the overall computation and data transmission [22, 29, 30, 33, 44, 45, 54, 59, 73]. Classic work such as fast R-CNN [29] use dedicated region proposal networks that are computationally heavy. Other approaches use simple extrapolation [22, 54, 73], which we find insufficient for eye tracking, because the objects (eyes) move rapidly. Many image sensors provide an ROI output mode [3, 4], but rely on users to provide the ROI coordinates. Sony built a sensor that automatically detects an ROI to drive spatial resolution and exposure modulation [47].

Our contribution is a lean and accurate ROI prediction algorithm tailored to eye tracking. We show that software-emulated events, combined with edge information from previous segmentation results, can effectively predict the eye movement and, by extension, the ROI.

3 EVENT-DRIVEN AUTO ROI FOR EYE TRACKING

Gaze estimation models rely on the geometry of foreground eye parts such as the pupil, iris, and sclera. Full-resolution eye images captured by near-eye cameras usually contain a large chunk of eye muscle/skin that is irrelevant to gaze tracking. We introduce an Auto ROI mode for eye tracking, where we predict and process only the ROI that contains the foreground eye classes needed for gaze estimation. This Auto ROI mode is coupled with a new and efficient eye segmentation network.

We first provide an overview of the algorithm (Section 3.1), followed by our feedback-driven, event-based ROI prediction algorithm (Section 3.2) and the eye segmentation network (Section 3.3).

3.1 Overview

Figure 2 shows a high-level flow diagram of our gaze tracking algorithm. We use a model-based, two-stage algorithm for gaze tracking. This paper’s contributions lie in the first stage, i.e., eye segmentation, while relying on a commonly-used gaze estimation model for the second stage [63]. Gaze estimation is relatively more mature and is much less compute intensive than the first stage (regression vs. DNN). Figure 2 thus omits the common gaze estimation part.

Initially at time T , the full-resolution eye frame is processed directly by the segmentation network to generate the segmentation map. The next frame at $(T + 1)$, instead of being processed directly,

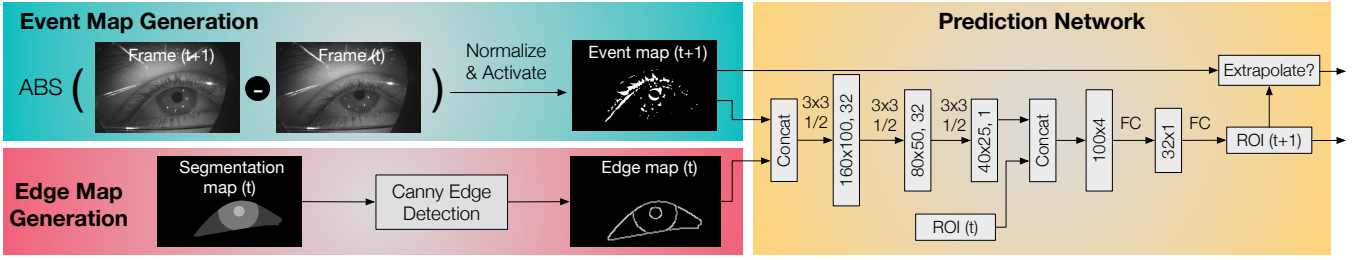


Fig. 3: The process of predicting the ROI at time $(T + 1)$ consists of three steps. First, we compute the absolute difference of frames at T and $(T + 1)$ to generate an event map. Second, we use Canny edge detection on the segmentation map at time T to extract an edge map. Finally, we concatenate the edge map and the event map to form the input to the ROI prediction network. The first three layers are Conv layers with 3×3 kernels followed by a Maxpool layer to reduce each dimension to $1/2$. The output of the last Conv layer is vectorized and concatenated with the ROI from time T (a 1×4 vector). The concatenated vector then goes through two FC layers to generate the predicted ROI of time $(T + 1)$. While our algorithm relies on events, the two additional cues of ROI and edge map from the previous frame are critical to the accuracy.

generates an event map, which is of the same dimension as the original full-resolution image. Each event map pixel is a 0/1 bitmask, indicating whether the corresponding pixel in the original image has a large change in intensity. This process emulates an actual event camera with simplifications tailored to eye segmentation.

The event map provides a useful guidance to predict the ROI in the current frame. We define the ROI as the minimal bounding box that encloses the three foreground eye parts, i.e., the pupil, iris, and sclera. Our approach can also apply to other tracking algorithms that use fewer or more segments (e.g., [46, 68]).

Events *alone*, however, are not robust enough. When the background eye muscles move significantly and/or when foreground eye parts move little between frames, activated events do not accurately capture the segment boundaries. To improve the robustness, we propose a feedback mechanism, where two important cues from time T (previous frame) are used to augment the ROI prediction.

For cases where the entire eye moves little between frames, e.g., at time $(T + 2)$ in Figure 2, our ROI prediction algorithm would detect the inactivities and opt to extrapolate from the previous segmentation map. Having this mode ensures “activity-proportional” tracking, where little tracking work is done when little activity is observed.

In most cases, only the first frame has to be processed in its full resolution. In rare ($\sim 0.02\%$) cases where the predicted ROI is physically infeasible (e.g., the top-right corner is to the left of the bottom-left corner), we fall back to the full-resolution mode.

3.2 Auto ROI with Event-Driven ROI Prediction

The ROI prediction algorithm has two roles: 1) predicting the ROI of the current eye frame, and 2) deciding whether the current eye frame requires going through a full-fledged segmentation algorithm or can be extrapolated from the previous segmentation map. Figure 3 shows the structure of our ROI prediction network.

3.2.1 The Prediction Algorithm

Intuition The goal of ROI prediction is to filter out background pixels (eye muscles, skin, and eyelids) while leaving only the foreground eye parts, i.e., the pupil, iris, and sclera. We use events to guide the ROI prediction. The intuition is that the background parts do not move as significantly as the foreground eye parts. Thus, activated pixels in the event map mostly correspond to the foreground eye parts and provide a useful guidance to the ROI prediction.

We generate the event map by emulating the high-level process of an event camera [27]. We first calculate the pixel-wise absolute difference, ΔF_{t+1} , between two consecutive frames, F_t and F_{t+1} . Next, each difference value, $\Delta F_t(x, y)$, is normalized by the corresponding pixel value in F_t . The normalized pixel values go through an activation function, Φ , which generates an event when the normalized

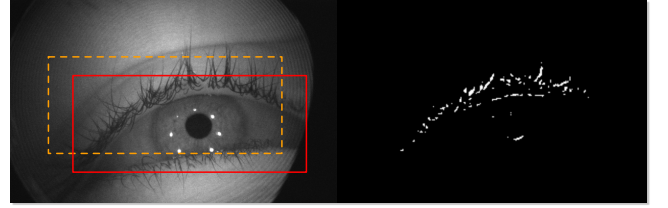


Fig. 4: Example of ROI misprediction using only the event map. In the left image, the solid ROI is ground truth, and the dashed ROI is the predicted ROI from using only the event map.

difference is greater than a predefined threshold value. Mathematically, generating an event map can be expressed as:

$$E_{t+1}(x, y) = \Phi(|F_t(x, y) - F_{t+1}(x, y)| / F_t(x, y), \sigma) \quad (1)$$

where $E_{t+1}(x, y)$ is the value at coordinate (x, y) of the event map at time $(T + 1)$, and Φ is the activation function which outputs 1 if the difference is greater than the threshold σ (and 0 otherwise). The threshold is a parameter that can be tuned for a specific application or scenario. Section 5.2 will show that the tracking accuracy is robust against the choice of σ , but there does exist a sweet spot. Normalizing pixel differences by the previous values mimics the log-scale absolute difference operation done by an actual event camera (i.e., $\log(a) - \log(b) = \log(a/b)$), where the pixel values are naturally in the log scale [27].

Two Feedback Cues While events are largely effective, there are cases where using events alone fails. In particular, when eye muscles/eyelids move significantly and/or foreground eye parts move little, events do not accurately capture the eye boundary anymore, challenging the assumption of using events to predict ROIs.

Figure 4 shows one such example, where the left panel shows the eye frame and the right panel shows the event map (generated from the current and previous frame). The dashed-line box is the predicted ROI using only the event map, whereas the solid-line box is the ground truth ROI. In this example the upper eyelid (irrelevant to gaze estimation) moves upward with the iris (used in gaze estimation); as a result, activated events capture both the eyelids and the iris, leading to a predicted ROI significantly above the actual eye.

To cope with the issue where events in the current event map do not accurately represent the foreground eye parts, we feed the previous ROI back to the prediction algorithm. The intuition is that the ROI of the previous frame provides useful information to predict the ROI in the current frame due to the temporal correlation. Thus, we feed the previous ROI into the ROI prediction algorithm (a DNN), which learns to correlate the previous ROI with the current ROI.

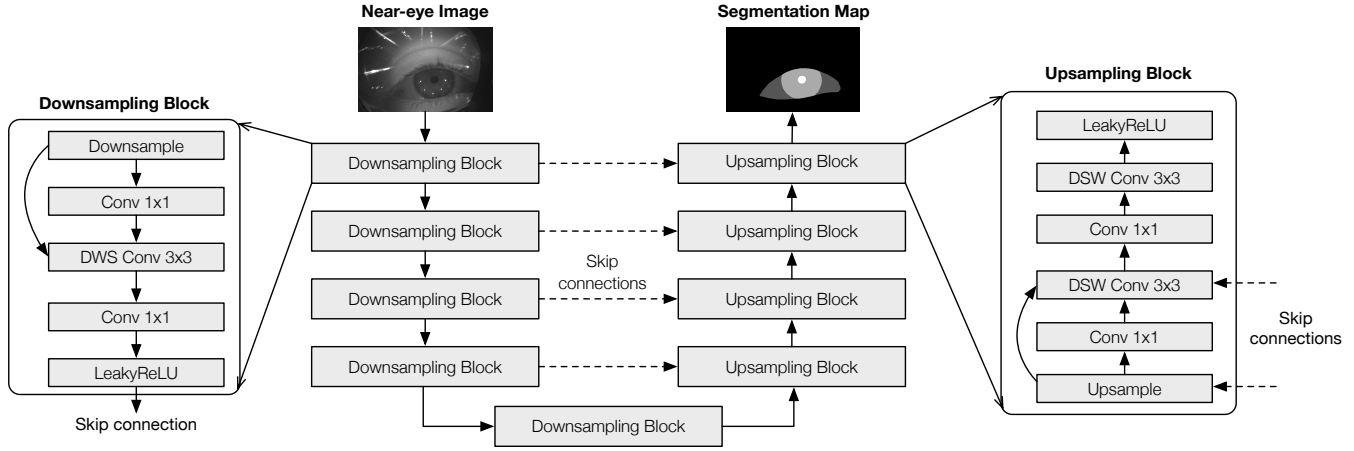


Fig. 5: The eye segmentation DNN. We use depthwise separable convolution as a building block to construct a U-Net-like architecture.

Using the previous predicted ROI, however, has an inherent downside: the ROI prediction errors will accumulate and the predicted ROI will drift over time. To mitigate error accumulation, we use an additional cue that does *not* drift over time — the previous segmentation map. The segmentation map is predicted directly from an actual eye image from the camera; thus, the segmentation map does not drift over time. To reduce data size, we extract an edge map from the segmentation map. The edge is defined as the boundary between two different classes in the segmentation map. We use Canny edge detection [13] to obtain an edge map from a segmentation map. Each pixel value in the edge map is a bitmask that indicates whether the corresponding pixel in the eye frame is a boundary.

Prediction Network With the guidance of the event map and the two feedback cues, ROI prediction can be very lightweight. Figure 3 shows the network architecture, which contains three convolution (Conv) layers and two full-connected (FC) layers. The event map and edge map are concatenated first and used as the input to the Conv layers due to the 2D nature of the two maps. In contrast, the ROI is a 1×4 vector; thus, it is concatenated with the flattened output of the Conv layers before entering the FC layers.

To reduce the overhead of the ROI prediction, we downsample the dimensions by 2 for both the event map and the edge map. We normalize the ROI to the image width/height so that each of the four coordinates in the ROI is within the $[0, 1]$ range. We find that it is easier for the network to learn normalized values, as opposed to the absolute coordinates, which vary by camera and scenarios.

3.2.2 Activity-Proportional Segmentation

In cases where the entire eyes do not move across frames, detecting the inactivities and skipping segmentation all together improves the execution speed. This allows our eye tracker to be activity-proportional: no work when no activity is detected. Two issues remain: how to *detect* inactivities and how to *compute* an accurate segmentation map extremely fast for inactive eye frames?

Building on top of our ROI prediction algorithm, we detect inactivity by calculating the event density of the event map inside the predicted ROI. If the event density is lower than a threshold γ , the current frame is deemed inactive, in which case the ROI prediction algorithm sets the extrapolation bit, indicating that no segmentation is to be executed for the current frame. While other inactivity detection schemes are possible, we favor the simple thresholding scheme because the threshold γ provides a useful knob to control the speed-vs-accuracy trade-off, allowing our system to potentially adapt to different application requirements and hardware capability.

We experiment with a range of extrapolation schemes, ranging from simply scaling the segments from the previous frame to using

a neural network to predict how to morph the segmentation from the previous frame. We eventually settled for the simplest scheme, where the previous segmentation result is reused. In retrospect, this scheme is the most robust because it relies the least on the inactivity detection scheme, which is a simple thresholding.

3.3 Efficient Eye Segmentation Network

The Auto ROI mode avoids full-resolution segmentation almost altogether, but the segmentation stage is still the single biggest bottleneck in the system, contributing to over 85% of the total execution time in our measurements. We propose a new segmentation network architecture that is much simpler without sacrificing accuracy.

Figure 5 shows our segmentation network inspired by U-Net [60]. Our architecture contains five downsampling blocks and four upsampling blocks, similar to the original U-Net design. However, unlike the original U-Net and prior eye segmentation networks [15, 46, 68] that rely only on convolutional layers (Conv), we judiciously use depthwise separable convolution (DWSConv) [17], a lightweight convolution primitive used in many efficiency-oriented networks such as MobileNet [35] to balance compute efficiency with accuracy.

We interleave 1×1 Conv layers with 3×3 DWSConv layers. The DWSConv layers are used to filter different channels independently, and 1×1 Conv layers are used to combine different channels' features. This interleaving leverages the efficiency of DWSConv while allowing channels to learn across each other to ensure accuracy. This strategy is used in both downsampling and upsampling blocks. We use skip connections to facilitate gradient flow in our network.

The convolution channel width, both in the Conv and the DWSConv layers, affects the network's ability to learn. We empirically find that DWSConv layers with skip connections usually require a higher channel width compared to other layers, because the learning abilities of those layers have global effects. As we will show, the channel width is an effective knob that provides a speed-vs-accuracy trade-off.

4 EVALUATION METHODOLOGY

4.1 Dataset and Ground Truth

OpenEDS 2020 We use the sequential data from the OpenEDS 2020 [55] dataset (Track-2)¹, which contains 200 sequences of continuous frames (29,476 frames in total) gathered from 152 participants with various ethnicities and iris colors. All sequences are captured by an infrared camera at 100 Hz with a 640×400 resolution. The average duration of a sequence is 30 seconds with diverse eye movements such as blinks and saccades (rapid eye movements). The

¹We do not use Track-1 data as the videos are only 1–2 seconds long.

average number of blinks per video is 4.5 (up to 18) and the average number of saccades per video is 7.8.

However, the sequential data in OpenEDS 2020 do not include the segmentation ground truth. To generate the ground truth for our evaluation, we train RITnet [15], a state-of-the-art eye segmentation network, on the non-sequential OpenEDS 2019 [28] dataset (100 Hz with a 400×640-pixel resolution), which does have ground truth segmentation labels. We then use the trained RITnet to generate the eye segmentation results for the sequential OpenEDS 2020 data.

The generated eye segmentation results are not perfect; we perform a series of data refinement steps to generate the final ground truth. We first apply the DBSCAN spatial clustering algorithm [21] on each initial segmentation result and identify the largest contiguous region as the correct eye region. We then fill the missing holes inside the eye region as a prior study does [43]. Finally, we manually inspected all sequences of the refined data and removed 15 sequences (out of 200), where the eye segmentations are visibly incorrect. In the end, our sequential dataset contains 185 sequences and 27,431 total frames.

With the eye segmentation ground truth, we then apply a gaze estimation algorithm [63] commonly used in prior work (e.g., DeepVOG [68]) to generate the gaze ground truth. Our algorithm does not depend on a particular gaze estimation method and, thus, can be integrated with other gaze estimation models [20]. We also use the eye segmentation ground truth to generate the ROI ground truth, which is the bounding box of the foreground segment (pupil, iris, and sclera). We manually verify that using so-defined ground-truth ROIs gives the same gaze results as those from full-frame inference.

TEyeD Another dataset we used in our evaluation is the TEyeD dataset, which has manually annotated segmentation and gaze ground truth [24]. We select 2 sequences from the dataset. Both sequences are grayscale images at a 384×288 resolution of captured at 25 Hz with an average length of over 30 minutes. In total, the selected data provide over 100,000 images.

On the TEyeD dataset, we report the 2D pupil position error (in pixels) rather than the 3D gaze direction error. This is because the gaze estimation algorithm [63] we use, not a contribution of this paper, does not provide accurate gazes for the TEyeD dataset.

4.2 Training

We first train a standalone eye segmentation network and a standalone ROI prediction network, and then refine eye segmentation using the ROI prediction results.

We follow prior studies and split the entire dataset into the training/validation/test sets with a 80/20 training/validation split, identical to EllSeg [46]. The eye segmentation network is trained using the Adam optimizer, a learning rate of 0.001, and a batch size of 4 for 250 epochs. We use the a loss function that combines both the standard cross-entropy loss and losses that are specialized to the eye structure [15]. To train the ROI prediction network, we use the Adam optimizer with a learning rate of 0.001, a momentum of 0.9, and the mean squared error loss function. We train the ROI prediction network for 100 epochs with a batch size of 8.

We then fine-tune the eye segmentation network using ROI images. In this step, the ROI prediction network first predicts an ROI given an eye image; we then crop the input image based on the predicted ROI and fine-tune the segmentation network with the cropped image. This training procedure is similar to networks based on region proposals such as Faster R-CNN [59]. We use a learning rate of 0.0001, a momentum of 0.9, and 100 epochs.

4.3 Baseline and Evaluation Metrics

We primarily compare against eye segmentation methods that are based on DNNs, which are shown to have superior accuracy compared to non-DNN-based segmentation methods [6, 25]. The baselines are trained using the same procedure as our algorithm.

- **RITnet** [15]: an encoder-decoder network using DenseNet [37] as the backbone. It won first place in the 2019 OpenEDS segmentation challenge [5]. As discussed in Section 4.1, we use RITnet to generate the ground truth for the sequential OpenEDS data. As a result, it is expected that our algorithm will have lower accuracy than RITnet.
- **DenseElNet** [46]: an ellipse segmentation framework for gaze tracking. DenseElNet is originally for ellipse segmentation (three segments). We re-purpose DenseElNet to predict four segments by modifying the channel size of the last layer.
- **DeepVOG** [68]: a popular encoder-decoder network for eye segmentation. DeepVOG was originally designed to generate only two segments (pupil and background). We modified the channel size of the last layer so that it predicts four segments.

We also compare the speed of our algorithms with the baselines. We measure the speed on a state-of-the-art mobile computing platform, Nvidia’s Jetson Xavier board [2]. We use the mobile Volta GPU on the Xavier board for all the networks. The GPU has 512 CUDA cores with a maximum frequency of 1,377 MHz.

4.4 Variants

We use two DNN variants of our eye segmentation network, Ours(S) and Ours(L). Both are designed with the same architecture, but differ in the channel width and thus provide a speed-vs-accuracy trade-off. In particular, Ours(S) has, on average, half the channels of Ours(L). Table 1 compares the amount of Floating Point Operations (FLOPs) and parameters of the two networks against the baseline segmentation networks.

Table 1: FLOPs and number of parameters in different eye segmentation networks for an input size of 640 × 400 8-bit pixels.

Network	DenseElNet	DeepVOG	RITnet	Ours (L)	Ours (S)
FLOPs (Billion)	53.1	36.5	23.1	2.6	1.2
Norm. FLOPs	45.2	31.1	19.7	2.3	1.0
# of Parameters (Thousand)	3047.3	2835.7	391.0	73.0	30.6
Norm. # of Parameters	99.6	92.6	12.8	2.4	1.0

The total computation requirement of Ours(L) is only about 1/9 of RITnet, 1/14 of DeepVOG, and 1/20 of DenseElNet. Ours(S) is even lighter weight, with about 45× fewer FLOPs and 100× fewer parameters compared to DenseElNet. Overall, Ours(S) uses only about 30K parameters (~120 KB).

To tease apart the contributions of the different components of our algorithm, we evaluate two variants on top of eye segmentation:

- **+ROI**: this variant uses only the ROI prediction (Section 3.2.1)
- **+ROI+E**: this variant uses both ROI prediction (Section 3.2.1) and extrapolation (Section 3.2.2)

5 EVALUATION

We first use OpenEDS 2020 to drive the analysis of the results. We present the overall performance and accuracy comparison (Section 5.1), followed by analyzing the sensitivity of our system to various parameters (Section 5.2) and algorithmic choices (Section 5.3). We perform an ablation study to evaluate the importance of the two feedback cues (Section 5.4). Finally, we present the results from the TEyeD dataset (Section 5.5) and analyze failure cases (Section 5.6).

5.1 Accuracy vs. Speed

Gaze Estimation Our algorithm achieves a 5.5× speedup over the baselines with a sub-0.5° gaze error. Figure 6a and Figure 6b compare the vertical and horizontal gaze errors and the speed of our algorithms with different baselines. The speedups are normalized to

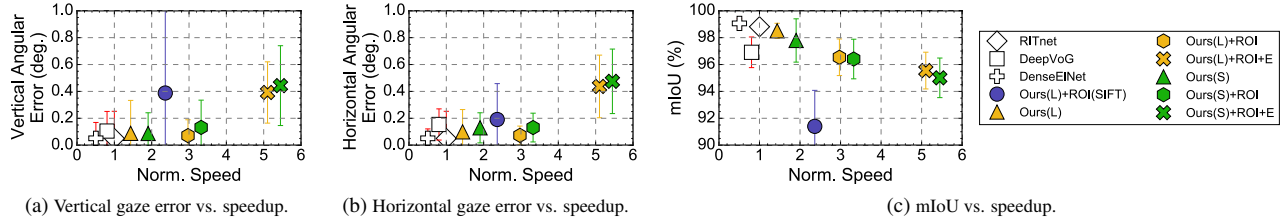


Fig. 6: The accuracy and speed comparison of different methods. All the subfigures share the same legend. The speedup values are normalized to the speed of RITnet. Ours(S) and Ours(L) are two eye segmentation networks. +ROI denotes ROI prediction is enabled. +E denotes the extrapolation is enabled. +ROI(SIFT) denotes using the SIFT-based ROI prediction.

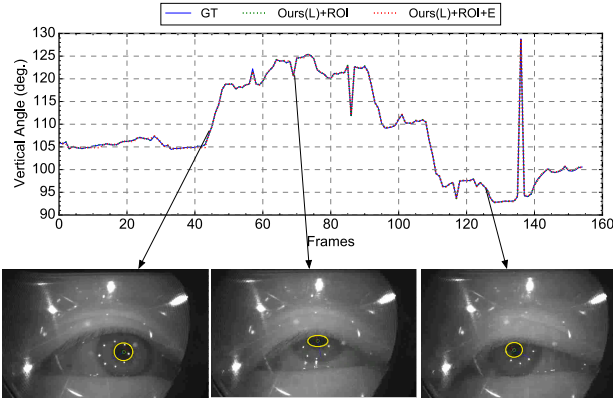


Fig. 7: Gaze estimation results over one sequence of frames. Ours(L) robustly tracks the ground truth. The bottom panel shows three representative cases: eye moves right, just before a blink, and eye moves up-left, respectively.

the speed of RITnet, which runs at 5.4 Hz on a mobile Volta GPU. Note that a 1° error is generally acceptable for gaze tracking [41].

RITnet, DenseEINet, Ours(L), and Ours(S) keep the absolute error rate below 0.1° in both the vertical and horizontal direction. They are all more accurate than DeepVOG. In particular, Ours(S) only has 0.04 and 0.05 higher gaze error than RITnet in each direction, but is $1.9\times$ faster. Note that this little gaze error loss is achieved with a network that is $12.8\times$ smaller (Table 1). This highlights our efficient eye segmentation network design (Section 3.3).

By operating on ROI images when possible, Ours(L)+ROI improves the speedup over RITnet to $3.0\times$. Interestingly, Ours(L)+ROI achieves better accuracy than Ours(L). Further inspection of the data shows that this is because by using only the ROI image for eye segmentation, we remove potential noise in the non-ROI region in the input eye image.

Using activity-proportional segmentation, Ours(L)+ROI+E and Ours(S)+ROI+E further improve the speedup to $4.2\times$ and $5.5\times$, respectively, while both retaining an angular error rate within 0.5° . The event density threshold used for extrapolation is set to 0.1%. We will study the algorithm’s sensitivity to this parameter in Section 5.2.

To further confirm the robustness of our system, Figure 7 compares the frame-by-frame gaze results of Ours(L), Ours(L)+ROI+E, and the ground truth. Ours(L) virtually matches the ground truth, and Ours(L)+ROI+E has slight deviations (e.g., around frame 10). We show three representative gazes in the bottom panel, where the eye moves right, blinks, and moves up left.

Finally, for a comprehensive analysis we show the gaze error distributions across all the evaluated frames in Figure 8. Not surprisingly, RITnet has the most compact distribution, because it is used to obtain the ground truth (see Section 4.1). Ours(L) has significantly better accuracy distribution compared to DeepVOG. Com-

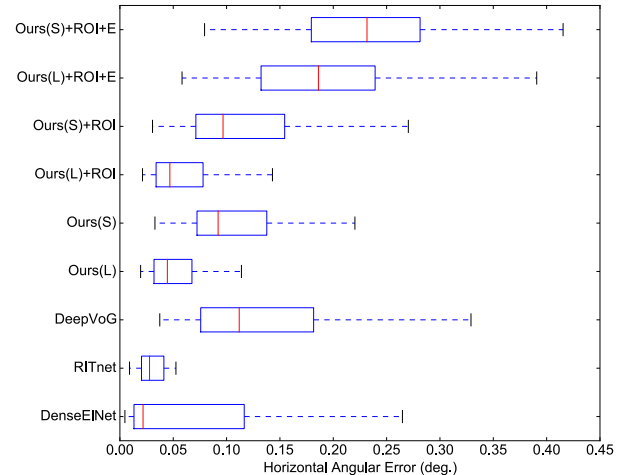


Fig. 8: Distributions of horizontal gaze error across as boxplots, which plot the median, 25th-percentile, 75th-percentile, the min, and max of the angular errors. RITnet is the most accurate because it is used to obtain the ground truth (see Section 4.1). Even the most inaccurate variant of our system, Ours(S)+ROI+E has a worst-case accuracy below 0.5° , which is generally regarded as an acceptable error bound for eye tracking [41].

paring to DenseEINet, Ours(L) has 0.02° lower average accuracy, but also has much better worst-case accuracy, as indicated by the significantly shorter tail. As we use a smaller network and exploit more speed-enhancing techniques, the error distribution generally shifts toward the right, but even the most inaccurate variant, Ours(S)+ROI+E, has a worst-case accuracy lower than 0.5° , which is regarded as an acceptable error bound for eye tracking [41].

Eye Segmentation While gaze accuracy is ultimately what we care about, we show the results of eye segmentation, an intermediate step to gaze estimation, to understand how the two metrics relate.

Figure 6c compares the eye segmentation accuracy and speed across different methods. DenseEINet achieves the highest mIoU at 99.0%. Ours(L) achieves a similar mIoU at 98.6%, but is $2.8\times$ faster and $20\times$ smaller. With a smaller network configuration, Ours(S) introduces a 1.2% mIoU loss. Ours(L)+ROI and Ours(S)+ROI have slightly lower mIoU at 96.5% and 96.4%, respectively. With extrapolation enabled, Ours(L)+ROI+E and Ours(S)+ROI+E can still keep the mIoU loss within 1.5%. Overall, our algorithm has slightly higher accuracy loss over RITnet on the segmentation metric than on the gaze error metric. This shows that the gaze estimation, a model fitting problem, can tolerate segmentation inaccuracy.

Frame Distribution On Ours(L)+ROI+E, 44.9% of frames perform extrapolations, 54.5% of frames are processed in the ROI mode,

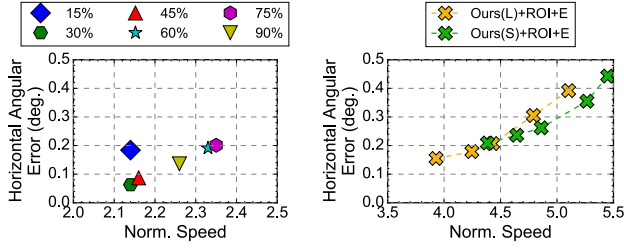


Fig. 9: Sensitivity of horizon- Fig. 10: Sensitivity of horizon-
tal gaze error and speedup to tal gaze error and speedup to dif-
ferent event activation thresh- ferent event density thresholds
olds σ . All results are from γ (from left to right: 0.001%,
Ours(L)+ROI. 0.005%, 0.01%, 0.05%, 0.1%).

and only 0.6% of frames require full-resolution eye segmentation. The vast majority of the frames that require full-resolution processing are the initial frames in each sequence. This frame distribution explains the speedup of our algorithm.

Time Distribution The time distribution of different stages of our algorithm points out opportunities for future optimizations. When executing Ours(L)+ROI+E on the mobile Volta GPU, the most time-consuming stage is the segmentation network, which takes 66.9% of total time. The second most time-consuming stage is the ROI prediction network, which takes 18.7% of the time. The event map and edge map are calculated on the CPU and take 3.6% and 3.9% of the time, respectively. The rest of the algorithm (data marshalling) takes 6.1% of the time.

We emphasize that speed we report in the paper should be seen as a *lower bound* of our algorithm and can be improved with more engineering efforts. For instance, the ROI prediction and eye segmentation networks can be pipelined across frames, and redundant memory copies can be removed if we directly manage the GPU memory rather than through PyTorch. Our goal here is to show that even unoptimized code (speed-wise) achieves several-fold speedup over state-of-the-art algorithms with virtually the same gaze accuracy.

5.2 Sensitivity Study

Event Activation Threshold Our algorithm emulates how an event camera generates events. The event activation threshold (σ in Equation 1) dictates how sensitive our algorithm is to eye activities. So far our evaluation has used a σ of 30% as the default value.

Figure 9 shows how σ affects the speed-vs-gaze trade-off, as σ increases from 15% to 90%. We show only the horizontal error; results for vertical angular error are similar and have been omitted. We retrain the ROI prediction network for each σ for a fair comparison. The speed is normalized to that of RITnet. The gaze errors under all σ settings are consistently lower than 0.5° , indicating an overall robust algorithm. 30% is an empirical sweep spot. A lower threshold, e.g., 15%, degrades accuracy, because a small threshold generates a noisy event map. Conversely, a higher threshold, e.g., 90%, has lower accuracy too, because a large threshold generates a sparse event map.

Event Density Threshold Figure 10 shows how the horizontal gaze error and speed vary with the event density threshold (γ), which we use to determine when to extrapolate (Section 3.2.2). γ is increased from 0.001% to 0.1% from left to right. As γ increases, more frames are extrapolated, so both the speed and the error increase. The gaze errors are consistently below 0.5° . Under a 0.001% γ , we achieve $3.9\times$ speedup over RITnet.

5.3 Design Decision Analysis

ROI Prediction Method To understand the effectiveness of our ROI prediction network, we construct a non-DNN method similar to those used in classic image stitching algorithms [10]. This method

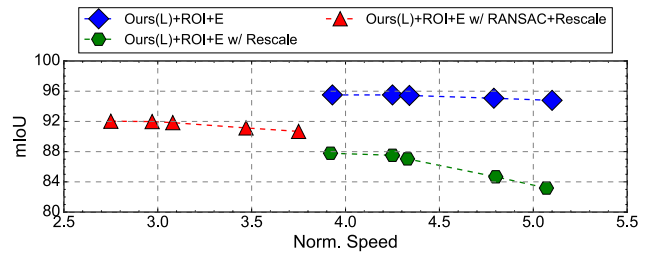


Fig. 11: Impact of extrapolation method on segmentation mIoU and speed normalized to RITnet. Extrapolation threshold values from left to right are 0.001%, 0.005%, 0.01%, 0.05%, 0.1%.

predicts the current frame’s ROI by matching SIFT features [49] from two consecutive image frames and using RANSAC [23] to calculate a transformation matrix between the two frames. Using the transformation matrix, we then compute the ROI coordinates of the current frame from the previous ROI. This variant is denoted Ours(L)+ROI(SIFT).

The gaze accuracy and the speed of Ours(L)+ROI(SIFT) are shown in Figure 6 in comparison to other methods. The gaze error of Ours(L)+ROI(SIFT) increases significantly, especially on the vertical direction, to 0.39 , much higher than Ours(L)+ROI (0.08). This suggests the effectiveness of our ROI prediction DNN.

Extrapolation Method The extrapolation method that we use in Section 3.2.2 simply reuses the previous segmentation map. To demonstrate its effectiveness, we compare with two other methods:

- **Rescale:** We first crop the previous segmentation map based on the previous ROI, and rescale the cropped segmentation map based on the size of the current, predicted ROI.
- **RANSAC+Rescale:** We use the ROI prediction network to first predict the ROI. The ROI is refined based on the SIFT features using RANSAC from the previous ROI. Lastly, we perform the same Rescale method as above.

Figure 11 compares the speedup (over RITnet) and segmentation accuracy of different extrapolation methods. All the extrapolation methods are applied to the Ours(L) network. We also vary the extrapolation threshold from 0.001% (left-most markers, fewest extrapolations) to 0.1% (right-most markers, most extrapolations).

We make two observations. First, our extrapolation method is consistently the most accurate. Combining RANSAC with rescaling is more accurate than simple rescaling, but also is much slower due to the overhead of RANSAC. Second, the accuracy generally drops with more extrapolations. But our method is the most robust to extrapolation. With 37.5% more extrapolations the mIoU degrades by only 0.7% (with $1.3\times$ speedup), whereas the accuracy under rescaling drops by 4.6%. The accuracy drop in rescaling-based methods is likely due to ROI error accumulation.

5.4 Ablation Study

The ROI prediction algorithm uses not only the event map, but also two feedback cues from the previous frame, i.e., the edge map and the ROI. Ablation studies here show the need for these two cues.

Previous ROI Figure 12a compares two ROI prediction networks: one only uses event maps and one uses event maps and previous ROIs. We train two ROI prediction networks using the same training procedure. We compare the mIoU and the speed of these two prediction schemes both with (+ROI+E) and without (+ROI) extrapolation. In the extrapolation mode, we sweep the event density threshold γ from 0.001% through 0.1%, same as in Figure 10. Without extrapolation, using the previous ROI improves the segmentation mIoU by about 1%, confirming that ROI feedbacks help predict future ROIs.

Table 2: Accuracy and speed on TEyeD dataset. The speed is normalized to RITnet. Both the segmentation accuracy (mIoU) and pupil position error in pixel (px) are shown with the standard deviation across the testset.

Network	RITnet	DeepVOG	DenseElNet	Ours (L)	Ours (S)	Ours (L)+ROI	Ours (S)+ROI	Ours (L)+ROI+E	Ours (S)+ROI+E
mIoU (%)	91.70 \pm 3.96	94.23 \pm 2.80	96.60 \pm 1.76	94.94 \pm 2.55	93.08 \pm 3.17	93.06 \pm 2.48	91.81 \pm 3.33	92.75 \pm 2.95	91.53 \pm 3.77
Pupil Error (px.)	0.55 \pm 0.17	0.49 \pm 0.17	0.28 \pm 0.09	0.42 \pm 0.13	0.50 \pm 0.18	0.46 \pm 0.13	0.62 \pm 0.25	0.49 \pm 0.10	0.65 \pm 0.23
Norm. Speedup	1	0.74	0.40	1.93	2.97	3.87	4.95	4.86	5.69

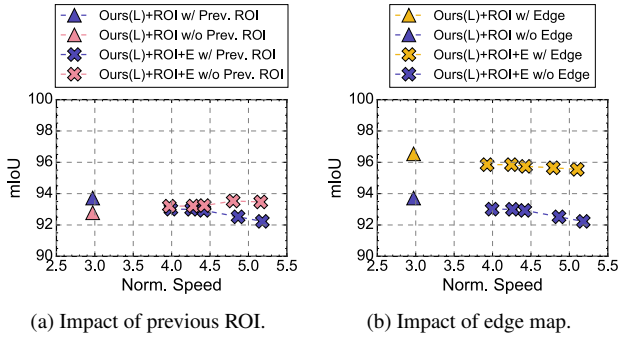


Fig. 12: Impact of using previous ROI and edge map in mIoU. Speeds are normalized to that of RITnet.

With extrapolation, the mIoU from using the previous ROI drops as γ increases. This shows how ROI prediction errors can accumulate. Interestingly, the mIoU without using the previous ROI *increases* as γ increases and is even higher than the mIoU without extrapolation at all. The reason, however, is that errors do not accumulate when previous ROIs are not used; when events are sparse (the prerequisite of extrapolation), the ROI naturally moves little, in which case using the previous ROI is the correct decision.

Edge Map Figure 12b compares two ROI prediction networks: one uses the edge map and the other does not, both trained with the same procedure. We show the segmentation accuracy and the speed with and without extrapolation enabled. Variants that use the edge map are consistently more accurate than ones that do not (2% mIoU difference). Using the edge map has little impact on speed, because the ROI prediction network is very lightweight to begin with.

Figure 13 uses a sequence of five consecutive frames to illustrate how the edge map helps correct the ROI drift. The solid and dashed rectangles are the ROIs that are predicted with and without using the edge map, respectively. The first frame is right after the user’s eye opens after a blink, in which case the event map provides a decent guidance to predict the ROI. However, in the next few frames the eye movements are insignificant, resulting in sparse event maps; in fact, the event map at frame 3 is completely empty. Thus, the ROIs predicted using the event map alone stay almost at the same location. The edge maps during the inactive period, however, still provide decent guidance as they are predicted from the segmentation maps.

5.5 Evaluation on TEyeD

Overall, the trend of the results on TEyeD matches that seen in OpenEDS 2020. Table 2 shows the accuracy and speedup across different methods. Compared to RITnet, Ours(S) achieves a 3.0 \times speedup with better accuracy; Ours(L) is the second best in overall accuracy, next to only DenseElNet, which is 4.8 \times slower and over 40 \times larger in model size (Table 1).

By incorporating Auto-ROI and extrapolation, Ours(S)+ROI and Ours(S)+ROI+E achieves a 5.0 \times and 5.7 \times speedup over RITnet, respectively, while still maintaining a sub-0.7 pixel pupil position error. The results show that our algorithm is robust on the large-scale TEyeD dataset, providing significant speedups over prior methods with higher or competitive accuracy.

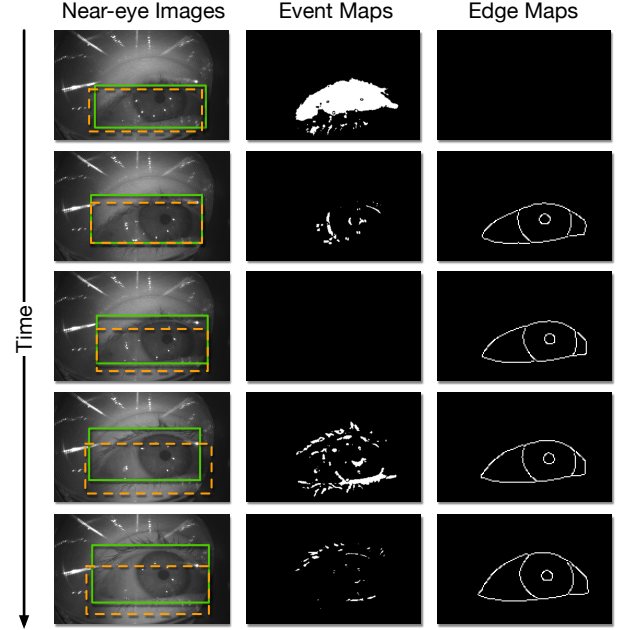


Fig. 13: A sequence showing that using the edge map helps correcting drifted ROIs. The solid and dashed rectangles are the ROIs predicted with and without using the edge map.

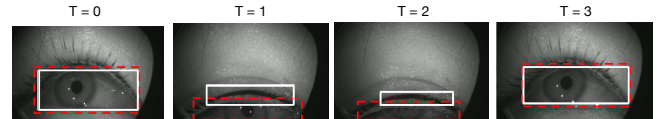


Fig. 14: A sequence showing failure cases. The solid and dashed boxes are the predicted and ground truth ROIs, respectively.

5.6 Failure Cases and Limitations

We show a failure case using a sequence of four consecutive frames in Figure 14. The solid and dashed boxes represent the predicted and ground truth ROIs, respectively. In just four frames, the user’s eye has significant vertical eye movements where the eye almost moves out of the camera view and then moves back. The predicted ROIs slightly drifted in the two middle frames. That said, the prediction model is robust enough to correctly predict the ROI after the eye moves back to the view center.

6 CONCLUSION

We present an algorithm that enables real-time eye tacking. The algorithm operates at real-time (over 30 Hz on a mobile GPU) — thanks to our Auto ROI mode, which judiciously processes only the ROI images whenever possible. Our main contribution is a novel ROI prediction algorithm, which emulates an event camera in software and uses events as natural guidance for ROI prediction. While events are largely effective, we show that feedback cues from previous frames are critical to ensure a robust ROI prediction.

REFERENCES

- [1] Non-dnn eye tracker. <https://github.com/YutaItoh/3D-Eye-Tracker>.
- [2] Nvidia Jetson AGX Xavier. <https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-agx-xavier/>.
- [3] Omnivision ov9782. <https://www.ovt.com/sensors/OV9782>.
- [4] Onsemi noip1sn025ka datasheet. <https://www.onsemi.com/pdf/datasheet/noip1sn025ka-d.pdf>.
- [5] Openeds 2019 challenge. <https://research.fb.com/programs/openeds-challenge/>.
- [6] C. Akinlar, H. K. Kucukkartal, and C. Topal. Accurate cnn-based pupil segmentation with an ellipse fit error regularization term. *Expert Systems with Applications*, p. 116004, 2021.
- [7] A. Al-Rahayfeh and M. Faezipour. Enhanced frame rate for real-time eye tracking using circular hough transform. In *2013 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1–6. IEEE, 2013.
- [8] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetstein. Event based, near eye gaze tracking beyond 10,000 hz. *arXiv preprint arXiv:2004.03577*, 2020.
- [9] M. Borys and M. Plechawska-Wójcik. Eye-tracking metrics in perception and visual attention research. *EJMT*, 3:11–23, 2017.
- [10] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, 2007.
- [11] T. T. Brunyé, T. Drew, D. L. Weaver, and J. G. Elmore. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive research: principles and implications*, 4(1):1–16, 2019.
- [12] M. Caligari, M. Godi, S. Guglielmetti, F. Franchignoni, and A. Nardone. Eye tracking communication devices in amyotrophic lateral sclerosis: impact on disability and quality of life. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14(7-8):546–552, 2013.
- [13] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.
- [14] S. Chandra, G. Sharma, S. Malhotra, D. Jha, and A. P. Mittal. Eye tracking based human computer interaction: Applications and their uses. In *2015 International Conference on Man and Machine Interfacing (MAMI)*, pp. 1–5. IEEE, 2015.
- [15] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz. Ritnet: Real-time semantic segmentation of the eye for gaze tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3698–3702. IEEE, 2019.
- [16] J. Chen and Q. Ji. 3d gaze estimation with a single camera without ir illumination. In *2008 19th International Conference on Pattern Recognition*, pp. 1–4. IEEE, 2008.
- [17] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.
- [18] V. Clay, P. König, and S. Koenig. Eye tracking in virtual reality. *Journal of Eye Movement Research*, 12(1), 2019.
- [19] G. Damian, T. Delbruck, and P. Lichtsteiner. Eye tracking using event-based silicon retina.
- [20] K. Dierkes, M. Kassner, and A. Bulling. A fast approach to refraction-aware eye-model fitting and gaze prediction. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pp. 1–9, 2019.
- [21] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, vol. 96, pp. 226–231, 1996.
- [22] Y. Feng, P. Whatmough, and Y. Zhu. Asv: Accelerated stereo vision system. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 643–656, 2019.
- [23] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [24] W. Fuhl, G. Kasneci, and E. Kasneci. Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. *arXiv preprint arXiv:2102.02115*, 2021.
- [25] W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci. Pupilnet: Convolutional neural networks for robust pupil detection. *arXiv preprint arXiv:1601.04902*, 2016.
- [26] W. Fuhl, T. C. Santini, T. Kübler, and E. Kasneci. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pp. 123–130, 2016.
- [27] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405*, 2019.
- [28] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019.
- [29] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [30] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.
- [31] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.
- [32] D. W. Hansen and A. E. Pece. Eye tracking in the wild. *Computer Vision and Image Understanding*, 98(1):155–181, 2005.
- [33] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [34] C. Hennessey, B. Nouredin, and P. Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, pp. 87–94, 2006.
- [35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [36] Z. Hu. Gaze analysis and prediction in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 543–544. IEEE, 2020.
- [37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- [38] Y. Itoh and G. Klinker. Interaction-free calibration for optical see-through head-mounted displays based on 3d eye localization. In *2014 IEEE symposium on 3d user interfaces (3dUI)*, pp. 75–82. IEEE, 2014.
- [39] R. J. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, pp. 573–605. Elsevier, 2003.
- [40] L. Jianfeng and L. Shigang. Eye-model-based gaze estimation by rgb-d camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 592–596, 2014.
- [41] A. Kar and P. Corcoran. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5:16495–16519, 2017.
- [42] H. Kim, S. Leutenegger, and A. J. Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision*, pp. 349–364. Springer, 2016.
- [43] S.-H. Kim, G.-S. Lee, H.-J. Yang, et al. Eye semantic segmentation with a lightweight model. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3694–3697. IEEE, 2019.
- [44] V. Kodukula, A. Shearer, V. Nguyen, S. Lingutla, Y. Liu, and R. LiKamWa. Rhythmic pixel regions: multi-resolution visual sensing system towards high-precision visual computing at low power. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 573–586, 2021.
- [45] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate

- region proposal generation and joint object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 845–853, 2016.
- [46] R. S. Kothari, A. K. Chaudhary, R. J. Bailey, J. B. Pelz, and G. J. Diaz. Ellseg: An ellipse segmentation framework for robust gaze tracking. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2757–2767, 2021.
- [47] O. Kumagai, A. Niwa, K. Hanzawa, H. Kato, S. Futami, T. Ohyama, T. Imoto, M. Nakamizo, H. Murakami, T. Nishino, et al. A 1/4-inch 3.9 mpixel low-power event-driven back-illuminated stacked cmos image sensor. In *2018 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 86–88. IEEE, 2018.
- [48] B. Li, H. Fu, D. Wen, and W. Lo. Etracker: A mobile gaze-tracking system with near-eye display based on a combined gaze-tracking algorithm. *Sensors*, 18(5):1626, 2018.
- [49] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157. Ieee, 1999.
- [50] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014.
- [51] K. Mania, A. McNamara, and A. Polychronakis. Gaze-aware displays and interaction. In *ACM SIGGRAPH 2021 Courses*, pp. 1–67. 2021.
- [52] P. Mathur, T. Mittal, and D. Manocha. Dynamic graph modeling of simultaneous eeg and eye-tracking data for reading task identification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1250–1254. IEEE, 2021.
- [53] M. L. Mele and S. Federici. Gaze and eye-tracking solutions for psychological research. *Cognitive processing*, 13(1):261–265, 2012.
- [54] B. A. Mudassar, P. Saha, Y. Long, M. F. Amir, E. Gebhardt, T. Na, J. H. Ko, M. Wolf, and S. Mukhopadhyay. A camera with brain-embedding machine learning in 3d sensors. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 680–685. IEEE, 2019.
- [55] C. Palmero, A. Sharma, K. Behrendt, K. Krishnakumar, O. V. Komogortsev, and S. S. Talathi. Openeds2020: open eyes dataset. *arXiv preprint arXiv:2005.03876*, 2020.
- [56] A. Patney, J. Kim, M. Salvi, A. Kaplanyan, C. Wyman, N. Bentley, A. Lefohn, and D. Luebke. Perceptually-based foveated virtual reality. In *ACM SIGGRAPH 2016 Emerging Technologies*, pp. 1–2. 2016.
- [57] A. Plopski, J. Orlosky, Y. Itoh, C. Nitschke, K. Kiyokawa, and G. Klinker. Automated spatial calibration of hmd systems with unconstrained eye-cameras. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 94–99. IEEE, 2016.
- [58] B. Ramesh, S. Zhang, Z. W. Lee, Z. Gao, G. Orchard, and C. Xiang. Long-term object tracking with a moving event camera. In *Bmvc*, p. 241, 2018.
- [59] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.
- [60] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.
- [61] T. Strandvall. Eye tracking in human-computer interaction and usability research. In *IFIP Conference on Human-Computer Interaction*, pp. 936–937. Springer, 2009.
- [62] L. Świrski, A. Bulling, and N. Dodgson. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the symposium on eye tracking research and applications*, pp. 173–176, 2012.
- [63] L. Świrski and N. A. Dodgson. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting [abstract]. In *Proceedings of ECEM 2013*, Aug. 2013.
- [64] Z. Wang, Y. Zhao, Y. Liu, and F. Lu. Edge-guided near-eye image analysis for head mounted displays. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 11–20. IEEE, 2021.
- [65] D. Weikersdorfer, R. Hoffmann, and J. Conradt. Simultaneous localization and mapping for event-based vision systems. In *International Conference on Computer Vision Systems*, pp. 133–142. Springer, 2013.
- [66] E. Whitmire, L. Trutoiu, R. Cavin, D. Perek, B. Scally, J. Phillips, and S. Patel. Eyecontact: scleral coil eye tracking for virtual reality. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, pp. 184–191, 2016.
- [67] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pp. 131–138, 2016.
- [68] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Oprey, V. L. Flanagan, P. Zu Eulenburg, and S.-A. Ahmadi. Deepvov: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of neuroscience methods*, 324:108307, 2019.
- [69] X. Zhang, X. Liu, S.-M. Yuan, and S.-F. Lin. Eye tracking based control system for natural human-computer interaction. *Computational intelligence and neuroscience*, 2017, 2017.
- [70] X. Zhang, Y. Sugano, and A. Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.
- [71] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4511–4520, 2015.
- [72] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.
- [73] Y. Zhu, A. Samajdar, M. Mattina, and P. Whatmough. Euphrates: Algorithm-soc co-design for low-power mobile continuous vision. *arXiv preprint arXiv:1803.11232*, 2018.
- [74] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 918–923. IEEE, 2005.