# Towards Mega-pixel Embodied Vision: Foveated Ego-View Predict Long-Horizon Contextual Intent in Dexterous Manipulation

**Yanbing Han[1], Ruilin Han[1,4], Gio Huh[1,5], Kevin Yang[1,6], Alan Yu[1], Jianan Wang[2] †, Ge Yang[1,3] †**

[1]FortyFive Labs, [2]Astribot, [3]MIT CSAIL, [4]Yale University, [5]Caltech, [6]Harvard University

†*Equal advising*

**Abstract:** Manipulating objects through physical contact on a moving robot demands perception that simultaneously captures the broad spatiotemporal context and task-critical visuomotor details. Biological vision addresses this tension through foveation: high-acuity sensing at the point of fixation, complemented by action-sensitive peripheral vision—an architecture that has independently evolved across species. In contrast, contemporary robotic systems typically downsample high-resolution camera streams, imposing a trade-off between field of view and visual detail that undermines decision-making. Our analysis demonstrates that human foveated active perception consistently predicts future task-relevant landmarks several hundred milliseconds in advance, providing long-horizon contextual cues for action planning while also revealing fine-grained details that indicate where visual attention should be directed. These findings suggest that wide-field passive vision systems today will be superseded by active perception that moves towards mega-pixel, foveated architectures.

**Keywords:** Eye tracking, Data Collection, Dexterous Manipulation

## 1 Introduction

Hand–eye coordination—the integration of actively controlled gaze with fine-motor hand movements—is fundamental to human motor skills. In competitive sports, for example, athletes enhance performance through eye-training exercises that deliberately regulate gaze to track projectiles or opponents' body cues. In everyday life, the absence of such coordination makes even simple pick-and-place tasks impossible. In contrast, robotic perception systems remain limited: they typically rely on passively mounted cameras with fixed intrinsics, where gaze control is replaced by soft-attention mechanisms computed from downsampled visual input. Despite the availability of modern cameras capable of streaming tens of millions of pixels per frame, robotic pipelines often operate only on VGA-resolution inputs, sacrificing visual detail for tractability.

This work addresses this gap by introducing a data-collection pipeline that synchronizes five critical streams: precise eye-gaze trajectories, high resolution egocentric RGB video, per-frame depth, real-time hand and head tracking. Our setup combines off-the-shelf extended-reality (XR) devices with targeted custom hardware, enabling fine-grained study of visuomotor coordination.

Our approach is motivated by cognitive and neural science research showing that gaze is not merely reactive but proactively guides motor behavior. In manipulation tasks, gaze trajectories consistently precede hand movements, thereby signaling future intent and reducing uncertainty [1]. Leveraging this principle, we seek to endow robots with the capacity to dynamically allocate sensing resources—zooming in on task-critical details while maintaining peripheral awareness of the broader environment.

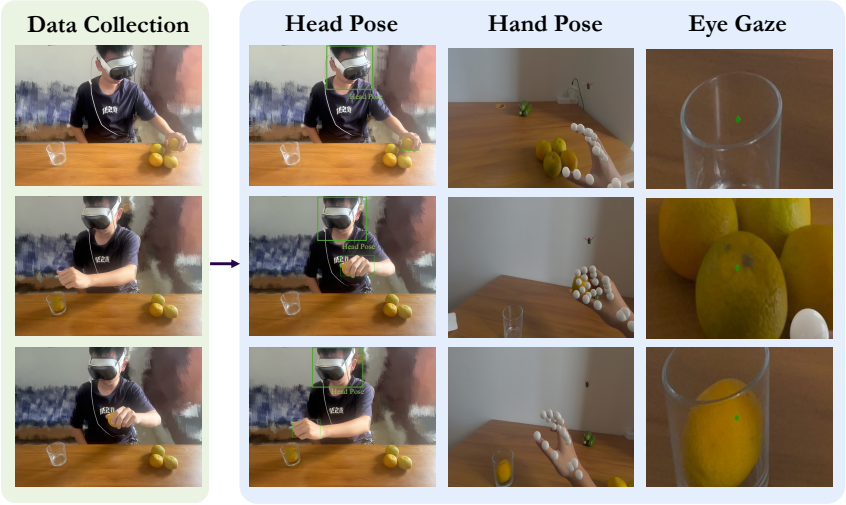


Figure 1: Humans coordinate head, hand, and eye movements to accomplish dexterous manipulation tasks. Our data pipeline captures this multimodal coordination by jointly recording eye gaze, RGB-D video, hand pose, and head pose using an extended-reality headset. By aligning these streams, we construct embodied datasets that couple human foveation with fine-grained motor trajectories. This allows for our further analysis of how gaze guides action, and provides rich supervision foundation for training robotic perception systems that anticipate and plan like humans.

At the same time, progress in robotics has been constrained by data limitations. Most existing datasets emphasize static image-label pairs or narrow action demonstrations, without high-resolution egocentric video with eye-tracking and hand motion. This restricts the development of embodied perception systems that can learn how visual attention modulates motor planning. Our protocol directly addresses this gap by capturing multimodal streams of foveated gaze, pixel-rich RGB-D video, hand trajectories, and head pose during dexterous manipulation (see Fig. 1).

Ultimately, our goal is to bridge the disconnect between human and robotic perception–action systems. Human motor control integrates peripheral monitoring with focal precision in a seamless perception–action loop, whereas robotic systems remain bottlenecked by passive visual pipelines and coarse attentional models. By grounding our dataset in the principles of biological foveation, we lay the foundation for next-generation robotic vision architectures capable of anticipating, planning, and executing complex actions in dynamic settings.

## 2 Methods

Modern extended-reality (XR) head-mounted devices are becoming increasingly common in everyday life, offering immersive experiences for work, learning, and entertainment. These systems generate rich streams of visual and proprioceptive data; however, access to such information is often constrained by privacy considerations [2]. In this work, we present a data collection pipeline for long-horizon dexterous manipulation that synchronously records high-resolution RGB image, depth image, human foveation signals, and hand and head poses using the Apple Vision Pro [3]. The Apple Vision Pro offers accurate, low-latency hand pose tracking, making it well-suited for fine-grained manipulation tasks. Nonetheless, its public APIs do not expose eye-tracking data or provide direct access to raw camera streams. To overcome these limitations, we augment the device with a custom eye-tracking algorithm and integrate an external camera system to capture both RGB and depth data.

### 2.1 From Pupil to Point of Regard: Preliminaries

Eye tracking objectively measures and records where a person looks and how their eyes move [4, 5]. This is achieved by measuring either the eye's position relative to the head or the orientation of the

gaze [4, 6]. Video-based eye tracking, which utilizes a camera directed at the eye to monitor its movements, has become increasingly popular in research due to its non-invasive nature and minimal impact on gaze behavior [4, 6]. Most remote and head-mounted eye trackers currently on the market employ video-based techniques, incorporating a camera alongside an infrared illumination source. In these systems, the pupil and cornea are essential for accurate tracking. Algorithms analyze the position of the pupil's center and the four corneal reflections relative to this center to determine the gaze point on a screen or within a scene, depending on the type of eye tracker used [4, 6]. The accuracy of this gaze mapping relies heavily on the precise measurement of these ocular features, allowing the system to track the eye's movements with high precision.

In our approach, we bypass explicit gaze-vector estimation and instead directly regress 2D screen coordinates from eye images. This process is followed by a light calibration step to align the outputs with the target display, ensuring accurate gaze mapping without the need for complex gaze-vector calculations.

### 2.2 Hardware setup and training data collection

**IR camera.** The Apple Vision Pro integrates two infrared (IR) cameras per eye, illuminated by lens-embedded IR LEDs, to enable precise gaze tracking. As the proprietary sensors are unavailable, we employ a low-cost infrared camera to capture eye images (Fig. 2).

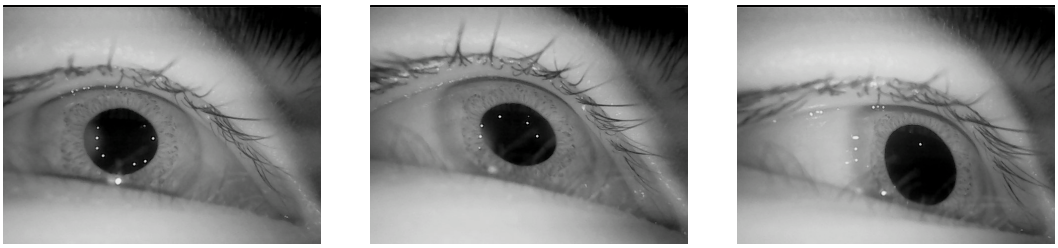

Figure 2: Infrared-illuminated eye images showing left, forward, and right gaze directions.

**Front view camera.** Since the front-view cameras are inaccessible, we employ a third-party RGB-D camera as an alternative. The camera is mounted onto the Apple Vision Pro, and the pose between the external camera and the headset is estimated. This pose enables us to project the predicted eye gaze into the image space captured by the mounted camera.

**Camera mount design.** We designed the front-view and infrared (IR) camera mounts and fabricated them with 3D printing. This approach ensures ease of reproduction and consistency in the resulting components.

**Training data collection.** We designed a controlled gaze-tracking experiment in which a single stimulus was presented on a full-screen $16 \times 16$ grid. The stimulus moved sequentially across grid cells, and participants were instructed to continuously fixate on it while corresponding eye images and gaze coordinates were recorded. Within each grid cell, the stimulus was positioned randomly to capture gaze targets across a continuous spatial distribution. Furthermore, the grid and point positions are defined relative to the head position, enabling data collection without requiring participants to remain stationary.

### 2.3 Model training for gaze estimation

For gaze estimation, we adopt a ResNet-50 backbone [7] as the feature extractor. The original classification head is removed and replaced with a regression head that directly predicts the 2D screen coordinates of the gaze point.

Formally, given an eye image $I$, the model $f_\theta$ predicts the normalized coordinates:

$$\hat{\mathbf{g}} = f_\theta(I), \quad \hat{\mathbf{g}} \in [-1, 1]^2 \tag{1}$$

where $\hat{\mathbf{g}} = (\hat{x}, \hat{y})$ denotes the predicted 2D screen coordinates.

To train the model, we employ the Huber loss [8], also known as the smooth $L_1$ loss, which balances sensitivity to small errors with robustness against outliers. For a prediction $\hat{\mathbf{g}}$ and ground truth $\mathbf{g}$, the point-wise Huber loss is defined as:

$$L_\delta(\hat{\mathbf{g}}, \mathbf{g}) = \begin{cases} \frac{1}{2}\|\hat{\mathbf{g}} - \mathbf{g}\|^2, & \|\hat{\mathbf{g}} - \mathbf{g}\| \le \delta, \\ \delta \cdot \|\hat{\mathbf{g}} - \mathbf{g}\| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \tag{2}$$

where $\delta$ controls the transition between the quadratic and linear regimes.

The training objective is the average Huber loss across all training samples:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} L_\delta(\hat{\mathbf{g}}_i, \mathbf{g}_i), \tag{3}$$

where $N$ is the number of training samples.

### 2.4 Stabilization, accuracy enhancements

We observe two primary failure modes in the baseline gaze regressor: (i) *temporal instability*, and (ii) *insufficient accuracy*.

**Temporal instability** refers to frame-to-frame jitter that arises even when the observer maintains fixation on a single point. It is caused by illumination flicker from the embedded LEDs, which produces moving light-stripe artifacts (Fig. 3), and by variations in pupil dilation (Fig. 4). To mitigate jitter while preserving responsiveness to genuine saccades, we introduce a *Transition Confidence Network (TCN)* that estimates the probability that a given pair of eye images corresponds to the same gaze target.

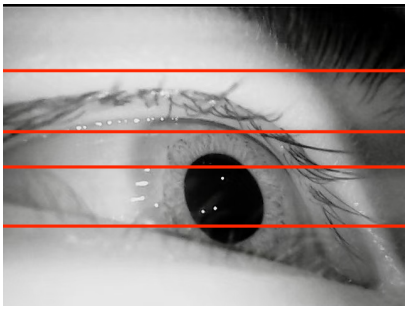

Figure 3: **Moving light stripe artifacts.** Illumination flicker from the embedded LEDs produces horizontal light stripe artifacts, highlighted by red horizontal lines, across the eye image.

We record extended fixation sequences on a single target, capturing diverse illumination stripe patterns and natural pupil dynamics for constructing data pairs. Each image in the pair is encoded independently using the same shared-weight encoder $E_\phi$ with a ResNet-18 [7] backbone. Specifically, given two images $I_1$ and $I_2$, their corresponding embeddings are obtained as follows:

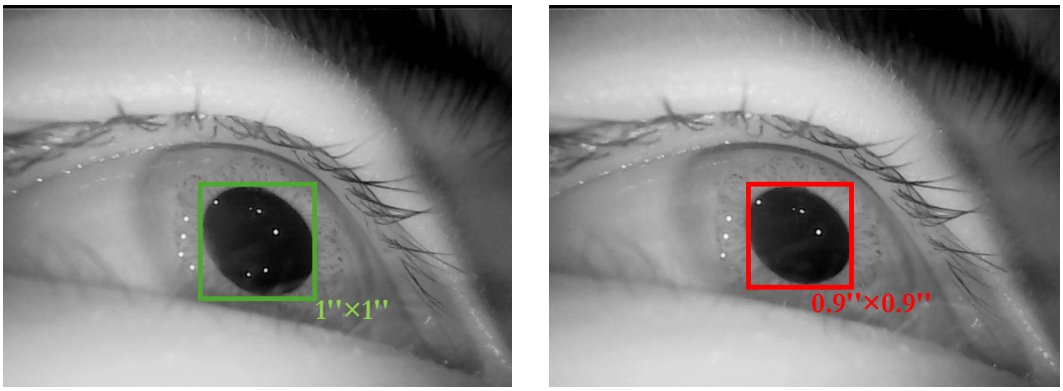


Figure 4: **Pupil dilation annotated with bounding boxes.** **Left:** Pupil at its larger state. **Right:** Same eye with a smaller pupil.

$$\mathbf{z}_1 = E_\phi(I_1), \mathbf{z}_2 = E_\phi(I_2) \tag{4}$$

These embeddings are then used to assess the similarity between the two images in the pair. We compute the predicted probability of the same-gaze:

$$p = \frac{1}{1 + e^{-(w\,\mathbf{z}_1^\top \mathbf{z}_2 + b)}} \tag{5}$$

where $w, b$ are learned scalars.

We train the TCN using the binary cross-entropy loss:

$$\mathcal{L}_{\text{TCN}} = -\left[c \log p + (1 - c) \log(1 - p)\right] \tag{6}$$

where $c \in \{0, 1\}$ denotes the ground-truth label.

At inference, the TCN output serves as a gating signal applied to the regressor's predictions. Specifically, let $\hat{\mathbf{g}}_t$ denotes the regressor's prediction at frame $t$ and $\tilde{\mathbf{g}}$ the stabilized estimate. Given a pair of successive eye images $(I_{t-1}, I_t)$, the TCN predicts the probability $p_t$ that the two images correspond to the same gaze. If $p_t \geq \gamma$, the stabilized output is preserved; otherwise, it is reset to the current regressor prediction:

$$\tilde{\mathbf{g}} = \begin{cases} \tilde{\mathbf{g}}, & \text{if } p_t \geq \gamma, \\ \hat{\mathbf{g}}_t, & \text{if } p_t < \gamma, \end{cases} \tag{7}$$

where $\gamma$ is a confidence threshold.

This strategy suppresses within-fixation jitter while allowing for rapid response at saccade onsets.

**Insufficient accuracy.** Each time the headset is donned, small pose changes introduce a session-dependent drift between the eye camera and display coordinate frames, degrading gaze prediction accuracy (Fig. 5). To compensate for this bias without retraining the model, we apply a lightweight post-hoc calibration to the network outputs.

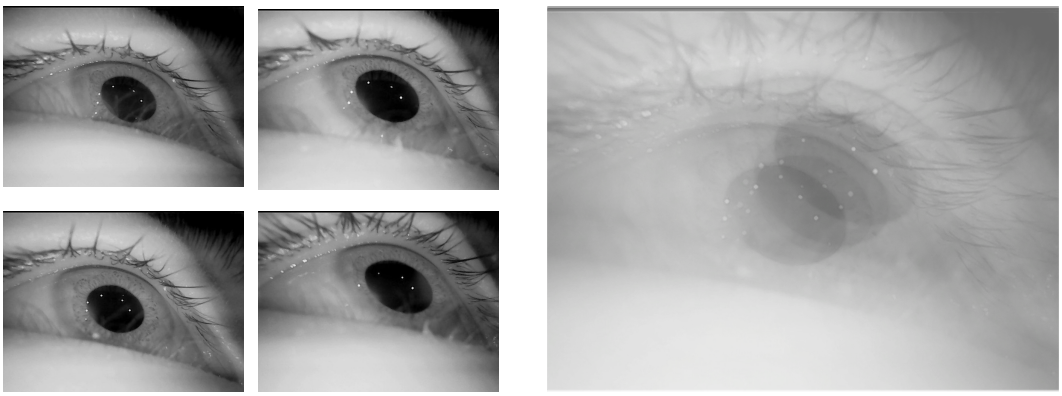

Figure 5: Session-dependent drift in gaze prediction. Left: eye images from different sessions with the same gaze direction. Right: the alignment of these images, highlighting session drift.

At the beginning of each session, participants sequentially fixate on eight on-screen targets. For each fixation $i$, we record the corresponding eye image $I_i$ and the ground-truth 2D gaze position $\mathbf{g}_i \in \mathbb{R}^2$. The network (Sec. 2.3) predicts the gaze position as $\hat{\mathbf{g}}_i = f_\theta(I_i)$. To account for session-specific variability, we estimate a transformation that best aligns the predicted positions with the ground truth by solving the following Procrustes problem:

$$\min_{s \in \mathbb{R}_+,\, \mathbf{t} \in \mathbb{R}^2} \sum_{i=1}^{8} \left\| \mathbf{g}_i - \left(s\,\hat{\mathbf{g}}_i + \mathbf{t}\right) \right\|_2^2. \tag{8}$$

The optimal parameters $(s^\star, \mathbf{t}^\star)$ obtained from this calibration are then applied to all subsequent predictions in the session:

$$\mathbf{g}^{\text{cal}} = s^\star \, \hat{\mathbf{g}} + \mathbf{t}^\star. \tag{9}$$

This calibration procedure effectively compensates for session-specific drift and enhances accuracy, while leaving the underlying network unchanged and incurring minimal computational overhead.

# 3 Applications

## 3.1 Collecting Robot-Free Real-World Dexterous Demonstrations

Using the Apple Vision Pro headset, we capture synchronized streams of hand and head poses, high-resolution RGB and depth images, and eye-gaze signals during real-world manipulation. These multimodal recordings provide rich embodied demonstrations that can be re-targeted to robots, enabling them to replicate human gestures and achieve more precise control in long-horizon dexterous manipulation tasks.

## 3.2 Annotating Gaze-Free Passive Videos

We enhance existing video datasets by annotating them with eye-gaze information. By replaying these videos in a VR environment, we can capture real-time gaze trajectories as users watch and interact with the content. This augments otherwise static video corpora with foveation signals, enriching them with human attention cues useful for robotics training. The resulting gaze-informed datasets enable robots to learn not only from controlled demonstrations but also from large-scale community video resources. This approach also has limitations (See Sec 4.3).

# 4 Results: Understanding Situational Eye-Gaze Data In Embodied Contexts

Our eyes are the window to the outside world, and a window to our inner world. They tell as much of a story about us as the rest of our bodies.

## 4.1 Reading Future Intent from Gaze Trajectories

Eye gaze reliably precedes motor actions, providing a powerful signal of future intent. To demonstrate this, we implemented a target-selection task in our Vision Pro APP. On each trial, four buttons were displayed, with one randomly designated as the target. Participants were instructed to select the target by clicking on it. We measured two key variables: (i) $T_{\text{gaze}}$, the first frame in which the gaze landed on the target, and (ii) $T_{\text{click}}$, the first frame in which the participant clicked the target. The gaze-to-click latency for each trial was then defined as $T_{\text{lag}} = T_{\text{click}} - T_{\text{gaze}}$.

Table 1 reports the mean gaze-to-click latency across tasks. On average, gaze preceded the click by approximately 790 ms, showing that gaze reliably reveals intent before action. This anticipatory property offers a valuable prior for robot learning, enabling earlier and more responsive goal inference.

| | **1** | **2** | **3** | **4** | **Average** |
|---|---|---|---|---|---|
| $T_{\text{lag}}$(ms) | 902.12 | 588.45 | 704.35 | 964.96 | 789.97 |

Table 1: Gaze-to-click latency, showing that eye gaze indicates future intent before action execution.

## 4.2 Active Foveation: A Key Driver of Precise Motor Skills and Hand–Eye Coordination

We conduct a randomized controlled experiment to test whether task-relevant gaze shifts ("active foveation") improve hand–eye coordination. Participants were split into two cohorts, either with or without structured gaze training. Task performance was measured by completion time-based success

| Task | w/o FOV | w/ FOV |
|---|---|---|
| *Easy* | | |
| Touch a button | <1 | <1 |
| Pick up a ball | <1 | <1 |
| Open a door handle | <1 | <1 |
| *Medium* | | |
| Pick up a cup | 2.1 | <1 |
| Stack one block on another | 1.3 | 1.02 |
| Open a drawer | <1 | <1 |
| Hand an object to someone | 1.4 | <1 |
| *Hard* | | |
| Pour water into a cup | 2.6 | 2.3 |
| Insert a key into a lock | 3.1 | 1.2 |
| Peel a banana | 1.8 | 1.9 |
| **Success Rate** | 0.7 | 0.9 |

Table 2: **Task completion time and success rate with and without active foveation.** Red indicates tasks that did not meet the expected time benchmark.

rate, where success is defined as completing the task within a baseline human-defined expected time threshold.

The gaze-trained cohort consistently outperformed the control group, completing tasks in less time and achieving higher success rates. Improvements were strongest in tasks requiring fine spatial alignment, reinforcing the role of active foveation in enhancing hand–eye coordination.

Eye-tracking overlays further illustrate the effect. Under the foveation condition, gaze trajectories were dispersed and purposeful, consistently targeting task-relevant features. In contrast, under the non-foveation condition, gaze remained largely fixed, resulting in overlapping pupil traces and lower task success.

### 4.3 Key Differences Between Passively Labeled and Actively Collected Behavior Data

Eye-gaze signals from passively labeled datasets are less predictive than those collected actively.

Suppose the task is *pick the orange into its category*. In passive labeling, gaze lacks contextual grounding, so it is difficult to infer intent until after an action begins. If both a cucumber and an orange are on the table, gaze alone does not reveal which object will be picked up. In this sense, passive gaze typically follows intention rather than predicting it. In active labeling, by contrast, gaze is directly anchored to the specified task, so it immediately reflects the user's intention and is therefore far more informative. Passive gaze signals can be improved by adding context before labeling (e.g., task prompt like "pick up the orange").

## 5 Conclusion

In conclusion, this work emphasizes the potential of foveated active perception to advance robotic systems, mimicking the human visuomotor coordination process. By introducing a novel data collection pipeline that synchronizes high-resolution RGB-D video, gaze trajectories, and hand and head motion, we lay the foundation for next-generation robotic vision architectures. Our findings suggest that this approach can improve task planning and execution in dynamic environments. In the future, we aim to explore the integration of this data into robotic systems, enabling robots to anticipate, plan, and execute complex actions with greater efficiency and context-awareness.

**Acknowledgments**

We would like to thank the anonymous reviewers for their insightful and constructive comments, which helped improve the clarity and quality of this work. We are also grateful to colleagues and peers who participated in valuable discussions and provided thoughtful feedback during the development of this research.

## References

[1] U. Sailer, J. Flanagan, and R. Johansson. Eye–hand coordination during learning of a novel visuomotor task. *The Journal of Neuroscience*, 25(39):8833–8842, 2005. URL https://www.jneurosci.org/content/25/39/8833.

[2] M. O. Rosenblat. Apple's data access limits on its vision pro are good for privacy – and also good for its business. https://bhr.stern.nyu.edu/quick-take/apples-data-access-limits-on-its-vision-pro-are-good-for-privacy-and-also-good-for-its-busir

[3] Apple Inc. Apple vision pro. https://www.apple.com/apple-vision-pro/.

[4] E. Kasneci, H. Gao, S. Ozdel, V. Maquiling, E. Thaqi, C. Lau, Y. Rong, G. Kasneci, and E. Bozkir. Introduction to eye tracking: A hands-on tutorial for students and practitioners, 2024. URL https://arxiv.org/abs/2404.15435.

[5] N. J. Wade and B. W. Tatler. *The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research*. Oxford University Press, 2005. ISBN 9780198566175.

[6] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer Publishing Company, Incorporated, 3rd edition, 2017. ISBN 3319578812.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[8] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. URL https://doi.org/10.1214/aoms/1177703732.