Democratic and People's Republic of Algeria
Ministry of Higher Education and Scientific Research
University Mohamed El Bachir El Ibrahimi of Bordj Bou Arreridj
Faculty of Mathematics and Computer Science
Department of Computer Science

FINAL DISSERTATION
Submitted as a Partial Fulfillment for the Degree of
Master in Computer Science
Speciality : Business intelligence engineering

TOPIC

Eye tracking in simple visual search tasks

Presented by :
BOUHADDA KENZA
BOUDIAF FELLA

Defended Publicly the : 12/06/2025
In front of the jury composed of:
President: Laifa Meriem
Examiner: Nouioua Mourad
Supervisor: REGOUID MERIEM

2024/2025

Dedication

First of all I want to thank me. I want to thank me for believing in me. I want to thank me for doing all this hard work.

Then to my family, thank you for your love, encouragement, and patience. You were my strength through every step of the way.

To my second family — Rahma,Rania,Aya,Souad,Meriem,Doua — and my Ouali thank you for being by my side through it all. You've always been there with laughs, advice, and motivation when I needed it most.

To my supportive friends across all the wilayas (Amira,Asma,Amel and M.Amira), your words, motivation, and good energy always reached me, no matter the distance.

To my work friends (Mouna,Halima..) you made the hard days easier and kept me motivated.

And a special thanks to my managers at work who always supported my decision to study while working and never stood in the way ,this meant the world to me and made me balancing everything possible.

This achievement is not just mine, but ours. I'm so grateful to all of you .

Kenza

Dedication

To my beloved family, Thank you for being my foundation, my strength, and my constant source of encouragement. Your unwavering love and support have carried me through every challenge and made every achievement meaningful.

A special dedication to my mother, whose sacrifices, patience, and unconditional love have shaped the person I am today. Your belief in me, even when I doubted myself, gave me the courage to persevere. This milestone is as much yours as it is mine. I love you more than words can say.

Fella

Acknowledgment

Above all, we would like to thank God for granting us the strength, perseverance, and wisdom throughout our journey in completing this thesis. Without His blessings and guidance, this accomplishment would not have been possible.

We express our sincere gratitude to our supervisor,**Dr. Regouid Meryem** , for granting us the opportunity to undertake this research and for her continuous and invaluable guidance throughout the entire research process.

We're also really thankful to our classmates and friends for all their help and support it made facing the challenges much easier.

Last but not least, we want to thank our families for their endless love, patience, and encouragement throughout this whole journey. Their faith in us kept us motivated every step of the way.

Abstract

Eye tracking has become an essential technique for understanding human visual attention and behavior across a wide range of fields. One of the core challenges in this domain is accurately predicting gaze during visual search tasks. As traditional models using handcrafted features often lack generalizability,Recent advances in deep learning offers a powerful alternatives by enabling data-driven learning of complex spatial patterns in visual attention.

This research introduces a deep learning-based eye-tracking system aimed at predicting visual attention through saliency maps, using the uEyes dataset which features a variety of image categories including desktop, mobile, web, and posters, along with corresponding human eye-tracking data . the system employs a U-Net convolutional neural network optimized for pixel-level saliency prediction. The model is trained and evaluated using a robust set of performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Kullback-Leibler Divergence (KLD), Correlation Coefficient (CC), Histogram Similarity (SIM), and Accuracy.

This system showcases the effectiveness of deep learning in modeling human visual behavior in visual search tasks.

Keywords: Eye tracking, visual attention, visual search, saliency prediction, deep learning, U-Net, uEyes dataset, gaze prediction, saliency maps, performance metrics

Résumé

L'oculométrie est devenue une technique essentielle pour comprendre l'attention et le comportement visuels humains dans de nombreux domaines. L'un des principaux défis dans ce domaine est de prédire avec précision le regard lors des tâches de recherche visuelle. Les modèles traditionnels utilisant des caractéristiques artisanales étant souvent peu généralisables, les avancées récentes en apprentissage profond offrent une alternative performante en permettant l'apprentissage, basé sur les données, de schémas spatiaux complexes de l'attention visuelle.

Cette recherche présente un système d'oculométrie basé sur l'apprentissage profond, visant à prédire l'attention visuelle grâce à des cartes de saillance. Ce système utilise la base de données uEyes, qui présente diverses catégories d'images, notamment celles des ordinateurs de bureau, des appareils mobiles, du web et des affiches, ainsi que les données d'oculométrie correspondantes. Le système utilise un réseau neuronal convolutif U-Net optimisé pour la prédiction de saillance au pixel près. Le modèle est entraîné et évalué à l'aide d'un ensemble robuste de mesures de performance, notamment l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (MSE), la divergence de Kullback-Leibler (KLD), le coefficient de corrélation (CC), la similarité d'histogramme (SIM) et la précision.

Ce système démontre l'efficacité de l'apprentissage profond dans la modélisation du comportement visuel humain dans les tâches de recherche visuelle.

Mots-clés : suivi oculaire, attention visuelle, recherche visuelle, prédiction de saillance, apprentissage profond, U-Net, Base de données uEyes, prédiction du regard, cartes de saillance, mesures de performance

ملخص

أصبح تتبع حركة العين تقنيةً أساسيةً لفهم الانتباه البصري البشري وسلوكه في مجموعة واسعة من المجالات. ومن التحديات الأساسية في هذا المجال التنبؤ بدقة بموقع النظرة أثناء مهام البحث البصري. ونظرًا لأن النماذج التقليدية التي تستخدم ميزاتٌ مُصممة يدويًا غالبًا ما تفتقر إلى إمكانية التعميم، فإن التطورات الحديثة في مجال التعلم العميق تُقدم بدائل قوية من خلال تمكين النماذج من التعلم القائم على البيانات للأنماط المكانية المعقدة في الانتباه البصري.

يُقدم هذا البحث نظامًا لتتبع حركة العين قائمًا على التعلم العميق، يهدف إلى التنبؤ بالانتباه البصري من خلال خرائط البروز البصري، باستخدام مجموعة بيانات يو آيز التي تضم مجموعة متنوعة من فئات الصور، بما في ذلك صور لواجهات سطح المكتب، والهواتف المحمولة، وصفحات الويب، والملصقات، إلى جانب بيانات تتبع حركة العين البشرية المقابلة. يستخدم النظام شبكةً عصبية تلافيفية من نوع يو-نت، وهي شبكة مصممة لتحسين التنبؤ بالبروز على مستوى البكسل. تم تدريب النموذج وتقييمه باستخدام مجموعة قوية من مقاييس الأداء، بما في ذلك متوسط الخطأ المطلق، ومتوسط الخطأ التربيعي، وتباعد كولباك-ليبلر، ومعامل الارتباط، وتشابه المدرج التكراري، والدقة.

يسلط هذا النظام الضوء على فعالية التعلم العميق في نمذجة السلوك البصري البشري في مهام البحث البصري.

الكلمات المفتاحية: تتبع حركة العين، الانتباه البصري، البحث البصري، خرائط البروز، التنبؤ بالنظرة، التعلم العميق، شبكة يو-نت، مجموعة بيانات يو آيز، المقاييس الإحصائية للأداء، الواجهة الرسومية.

Contents

List of Figures

List of Tables

List of Equations

Chapter 1

General introduction

1.1 Context

In recent years, eye tracking has emerged as a valuable tool for understanding visual attention and human behavior by analyzing how individuals interact with visual stimuli such as images, interfaces, and advertisements across various fields like psychology, marketing, and healthcare. Traditionally dependent on expensive, specialized hardware, eye-tracking has become more accessible due to advances in computer vision and deep learning, which enable the prediction of gaze patterns using machine learning models based solely on image content. Leveraging these developments, This project focuses on creating a deep learning-based eye-tracking system that generates saliency maps to predict visual attention in static images. The system aims to replicate human gaze behavior and support visual search analysis by highlighting the most attention-attracting areas within an image.

1.2 Goals

The primary objective of this research is to develop and implement a deep learning based eye tracking system capable of accurately predicting saliency maps that reflect human visual attention on static images, while providing an intuitive graphical user interface (GUI) for users to upload images and view prediction results. The system's performance will be rigorously evaluated using established saliency metrics such as Mean Absolute Error (MAE),Mean Squared Error (MSE), Kullback-Leibler Divergence (KLD), Correlation Coefficient (CC), Similarity

(SIM), and classification based Accuracy. Additionally, the research aims to enhance understanding of visual attention patterns across various image types, including posters, mobile layouts, desktop screenshots, and web interfaces, thereby supporting diverse applications in visual analysis.

1.3 Methodology and results

To achieve these goals, the system was developed using a convolutional neural network (CNN) based on the U-Net architecture and trained on the publicly available uEyes dataset. The development process involved several key steps: first, data preprocessing where input images and corresponding saliency maps were resized to a uniform resolution, normalized, and converted to RGB format to ensure consistency.Second, model development utilizing TensorFlow to train the U-Net model to learn the mapping from images to human saliency maps. Third, interface design employing Gradio to create a user-friendly GUI for image upload, prediction visualization, and result interpretation. and finally, The system was tested using unseen validation data and evaluated with multiple metrics.

1.4 Report outline

This report is structured into six chapters, each addressing key aspects of the proposed eye-tracking system and its broader context:

Chapter 2 :focuses on eye tracking and visual search tasks, examining how eye movements are used to locate targets within complex scenes. It distinguishes between feature search,, and conjunction search. It also discusses the valuable role of eye-tracking technology in understanding visual exploration across various fields. Finally, it acknowledges challenges in visual search research.

Chapter 3 :provides a detailed overview of the fundamentals of eye tracking by explaining the anatomy of the eye, and describes key types of eye movements and their roles in visual perception and attention. The chapter also introduces important eye tracking metrics ,establishing the theoretical foundation necessary for interpreting eye tracking data and understanding how these physiological and behavioral components inform eye tracking research and applications.

Chapter 4: System Design and Implementation details the comprehensive development process of the eye-tracking system, covering data preprocessing and cleaning using the uEyes dataset, the design and training of the U-Net architecture for saliency map prediction, evaluation based on multiple saliency metrics, and visualization methods.

Chapter 5: Results outlines the technical setup and software environment employed for implementation, presents the Gradio-based graphical user interface (GUI) developed for user interaction, and discusses practical outcomes including saliency prediction outputs, evaluation metrics, and insights into model behavior.

Chapter 6: Conclusion and Future Work summarizes the project's achievements, reflects on the system's strengths and limitations, and suggests potential improvements and directions for further research.

Chapter 2

Eye Tracking and Visual Search Tasks

2.1 Introduction

Visual search tasks play a central role in understanding how people interact with complex visual environments. This chapter introduces the concept of visual search, defines its key characteristics, and explains how different types of searches, such as feature search and conjunctive search, affect visual behavior. The role of visual attention in guiding eye movements during search tasks is discussed, highlighting how eye tracking technology provides critical insights into where and how people focus their gaze. We also explore practical applications of eye tracking in fields like usability testing, healthcare, and security, while addressing common challenges researchers face in visual search studies. Together, these topics set the stage for a deeper understanding of how eye tracking and visual search are closely connected.

2.2 Definition of Visual Search Tasks

Visual search tasks involve scanning a scene with the eyes to either locate a specific target or confirm its absence. Unlike reading, where eye movements follow a structured and predictable pattern, visual search is more variable and depends heavily on the nature of the task, the type of image, and its content. During visual search, individuals make fixations and saccades to gather detailed information through foveal vision, which the brain then integrates into a coherent understanding of the scene. This process combines both serial (step-by-step) and parallel (broad) strategies and is influenced by visual features such as color, size, and shape [1].

2.3 Types of Visual Search Tasks

Visual search tasks are typically divided into two main types: feature search and conjunction search.

2.3.1 Feature search

Feature search refers to a type of visual search task in which the target is defined by a single, basic visual attribute such as color, size, orientation, or motion that makes it distinct from all surrounding distractors. For example identifying a vertical line among horizontal lines would be considered a feature search.These targets typically pop out from the visual scene and are detected quickly and effortlessly, regardless of how many other items are present [2].This efficiency is due to parallel processing mechanisms that operate in the early stages of visual perception.

According to Feature Integration Theory, basic features are processed automatically and simultaneously across the entire visual field without the need for focused attention. Because of this, the time required to locate a target in a feature search tends to remain stable even as the number of distractors increases, suggesting that attention is not required to scan each item individually.

Feature search is an example of preattentive processing, where simple features are rapidly and unconsciously analyzed before attention is directed toward any specific object. As a result, feature searches are generally fast, highly accurate, and require minimal cognitive effort, making them fundamentally different from more attention-demanding search types like conjunction search [1].

2.3.2 Conjunction search

Conjunction search involves locating a target defined by a unique combination of two or more features, such as a red circle among red squares and blue circles, and typically requires focused, deliberate attention. Because the target cannot be identified based on any single feature alone, observers must bind multiple features together to distinguish it from distractors.

This process is generally thought to occur in a serial manner, scanning items one by one,

as reflected in characteristic eye movement patterns and scan paths. According to Feature Integration Theory (FIT), while individual features are processed in parallel, attention is needed to bind them into a coherent percept, making conjunction searches slower and more attentionally demanding. However, Wolfe's Guided Search theory offers a more dynamic view, suggesting that early feature information can guide attention toward likely target locations, allowing the search to be more efficient than purely serial scanning [1].

2.4 Visual Attention

Visual attention plays a crucial role in guiding eye movements during visual search, directing our gaze toward the most relevant areas for detailed processing. This tight coupling between attention and eye movements allows us to efficiently interpret and interact with our visual environment [2]. As we move our eyes to bring specific regions of the visual field into high-resolution foveal vision, we actively seek detailed information, highlighting the connection between gaze shifts and the need for focused inspection.

Traditionally, visual attention has been framed through the distinction between the "where" spatial location, linked to peripheral vision and guiding eye movements and the "what" object identity and meaning, associated with foveal vision and detailed analysis. Eye movements provide observable (overt) evidence of attention, and scan paths reveal the temporal sequence of visual focus.

The control of visual attention is shaped by both bottom-up and top-down processes. Bottom-up attention is stimulus-driven, triggered by visually salient features such as color contrasts, edges, or motion, often resulting in the "pop-out" effect where attention is captured automatically. In contrast, top-down attention is guided by the viewer's goals, expectations, and prior knowledge meaning that eye movements are strategically directed to task relevant objects [3].For example, searching for a clock versus a microwave will produce distinct eye movement patterns.

To quantify how attention influences gaze behavior, eye-tracking metrics are commonly used, including search efficiency (time to fixate the target), selectivity (likelihood of fixating relevant items), and capture (probability of fixating salient but irrelevant distractors). Although it is generally assumed that attention aligns with foveal vision, research continues to explore

the nuances of this relationship [2].

2.5 Role of Eye Tracking in Visual Search

Eye tracking has become a widely used method for studying visual search, as it provides a direct and continuous measure of overt visual attention. Unlike traditional reaction time measures, eye tracking captures where participants look, how long they fixate, and in what sequence offering rich insights into the spatiotemporal dynamics of visual search behavior. By analyzing eye movements such as fixations and saccades, researchers can determine which objects receive attention, how long they are examined, and how the gaze moves across the scene.

This level of detail enables the assessment of key aspects of visual search, including efficiency, selectivity, attentional capture, and the influence of memory and prior knowledge. Eye-tracking studies span a wide range from controlled experiments using simple, abstract arrays to more naturalistic investigations involving complex, real-world scenes [2]. Ultimately, eye tracking serves as a powerful tool for uncovering both the visual and cognitive mechanisms underlying the active process of search [3].

2.6 Applications of Eye Tracking

Eye tracking is used across a wide range of fields, including:

2.6.1 Usability Testing UX Research

Designers use eye tracking to see where users look first and what draws their attention in digital interfaces or product packaging. It reveals which features are noticed or ignored, helping optimize layouts for better user experience.for example, ensuring a call-to-action button stands out [1].

2.6.2 Neuroscience Psychology

Eye tracking provides insights into how people read, search, and focus attention. It helps researchers study cognitive strategies and detect attention disorders like ADHD. Combined with

tools like fMRI or EEG, it links gaze patterns to brain activity, deepening our understanding of perception and decision-making [1].

2.6.3 Marketing Advertising

Marketers use eye tracking to analyze how consumers engage with ads. It shows which elements (logos, images, text) capture attention, guiding better design and placement strategies to improve recall and influence buying decisions [1].

2.6.4 Automotive Aviation

In cars and planes, eye tracking supports safety and training. Driver monitoring systems detect fatigue or distraction, while pilot training uses gaze tracking to ensure proper scanning of cockpit instruments. This reduces human error and enhances situational awareness [1].

2.6.5 Education

In educational settings, eye tracking is used to explore how students visually interact with learning materials, providing valuable data on their information processing strategies. By observing where students focus their gaze and for how long, educators can identify which parts of a text, diagram, or multimedia content are most engaging or challenging. This knowledge helps in designing teaching tools and curricula that align better with natural visual attention patterns, improving comprehension and retention. Eye tracking can also support personalized learning by detecting when students struggle to process information, allowing timely intervention [4].

2.6.6 Forensics

Forensic scientists and legal professionals use eye tracking to investigate eyewitness memory accuracy, deception, and other courtroom-related issues. By analyzing where and how witnesses look during recall or interrogation, researchers can gather objective evidence about their attention and cognitive state. Eye movement patterns can reveal signs of lying or uncertainty, supplementing traditional methods of lie detection. Additionally, eye tracking helps improve the reliability of eyewitness testimony by understanding how visual attention influences memory formation and recall, ultimately supporting fairer judicial outcomes [5].

2.7 Challenges in Visual Search Studies

While visual search studies offer valuable insights into attention and perception, they face several methodological challenges. A major difficulty lies in designing stimuli that strike a balance between naturalism and experimental control, especially when using complex, real world scenes [2]. Researchers also struggle to clearly separate the contributions of overt eye movements from covert attention the mental focus that may not always be accompanied by gaze shifts. Another challenge stems from the diversity of visual search tasks, which makes it difficult to generalize findings across different experimental contexts. Developing paradigms that can effectively isolate specific cognitive processes, such as the role of contextual information in object recognition over time, remains a significant hurdle. Additionally, with the rise of computational modeling, particularly deep learning approaches, the field increasingly requires large, well-annotated datasets to train models that accurately reflect human visual search behavior [6].

2.8 Limitations of Eye Tracking

Participant-related limitations in eye tracking often result from demographic, physiological, and cognitive differences, such as age, nationality, health status, and physical traits. These factors can affect the accuracy, reliability, and generalizability of eye-tracking data.

Age-Related Factors: Older adults may have reduced pupil size or slower eye movements, affecting calibration. Young children may struggle with focus and following instructions .

Cultural and National Differences: Reading direction, visual habits, and attentional strategies vary across cultures, influencing gaze behavior and limiting generalizability .

Health Conditions: Visual impairments (e.g., cataracts) and neurological disorders (e.g., Parkinson's, autism) can disrupt calibration and produce atypical eye movements .

Physical and Physiological Traits: Eye color, eyelid shape, and use of glasses or contact lenses can interfere with eye tracker accuracy, especially in infrared-based systems .

Participant Compliance and Fatigue: Inattention, tiredness, or misunderstanding tasks can introduce variability and reduce data reliability .

These individual-level factors highlight the importance of careful participant selection, adaptive study design, and contextual data interpretation to ensure that eye tracking research yields valid and generalizable insights [1].

2.9 Conclusion

In summary, this chapter provided an overview of visual search tasks and the critical role of attention and eye tracking in studying visual behavior. By examining different types of search tasks, key applications, and ongoing challenges, we build a foundation for understanding how visual information is processed and how future research can further uncover the complexities of human visual exploration.

Chapter 3

Eye Tracking Fundamentals

3.1 Introduction

Eye tracking has become a powerful tool for understanding human visual attention, perception, and behavior. This chapter provides a comprehensive overview of eye tracking fundamentals, starting with the eye anatomy and the eye tracking basic definition and the different types of eye movements, such as fixations, saccades, smooth pursuits, and blinks. It also traces the historical development of eye tracking technologies, from early mechanical methods to modern infrared and wearable systems. Key metrics and analysis methods like fixation duration, saccade amplitude, scanpaths, and heatmaps are discussed, alongside a review of important research contributions and existing eye tracking databases. Together, these topics lay the foundation for understanding how eye tracking has evolved and how it is applied in various fields today.

3.2 Eye anatomy

The human eye is a complex sensory organ responsible for vision, composed of three main layers or coats. The outermost layer, the fibrous tunic, includes the transparent cornea at the front and the white sclera that forms the majority of the eye's outer surface, providing structure and protection. The middle layer, called the vascular tunic or uvea, contains the choroid, ciliary body, and iris; this layer supplies blood to the eye and controls light entry via the iris, which adjusts the pupil size. The innermost layer is the retina, a light-sensitive tissue that converts

light into electrical signals sent to the brain through the optic nerve. The eye's interior is divided into chambers filled with fluids: the anterior chamber (between cornea and iris) and posterior chamber (between iris and lens) contain aqueous humor, which nourishes the eye, while the larger vitreous chamber behind the lens contains the jelly-like vitreous body that maintains eye shape and holds the retina in place. The lens, suspended by zonule fibers from the ciliary body, changes shape to focus light on the retina, enabling accommodation for near and far vision. Surrounding the eyeball are six extraocular muscles that control eye movement within the orbit. This intricate anatomy allows the eye to capture and process visual information effectively [07].

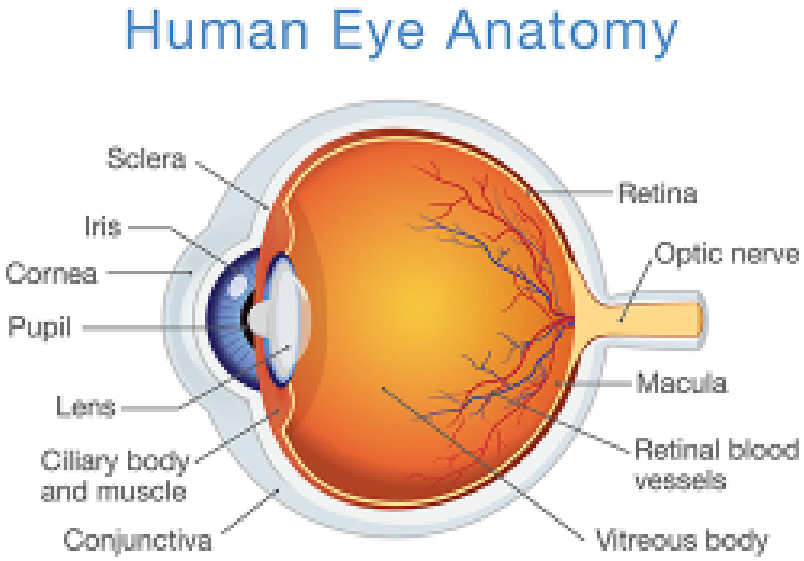


Figure 3.1: Human eye anatomy [08]

3.3 Definition of Eye Tracking

Eye tracking is an advanced technology and research method used to measure and analyze eye movements and gaze behavior. It captures where an individual is looking (gaze point), how their eyes move across a scene (saccades), and how long they fixate on specific areas (fixations). These data provide valuable insights into visual attention, cognitive processing, and behavioral responses during interaction with visual stimuli such as images, videos, websites, or real-world environments. This technology typically uses infrared light and high resolution

cameras to detect and record eye position and movement with precision. This information is then processed to identify patterns, including fixation duration, saccadic trajectories, and scan paths the sequential flow of eye movements across a visual field. Additionally, pupil dilation can be measured, offering further information about cognitive load or emotional arousal [1].

3.4 Eye Movement Types

Visual information is acquired through various types of eye movements, each contributing uniquely to perception and attention. Below are the main categories of eye movements:

3.4.1 Fixations

Fixations are moments of relative eye stillness during which the visual system processes information in detail. Although the eyes appear stationary, small involuntary movements such as microsaccades, drift, and tremor occur to maintain visual acuity and prevent visual adaptation [3].

3.4.2 Saccades

Saccades are rapid, jerky eye movements that occur when shifting attention from one point to another. These movements reposition the fovea (the area of the retina responsible for sharp vision) toward a new target. During saccades, visual processing is temporarily suppressed to prevent motion blur, a phenomenon known as saccadic suppression [3].

3.4.3 Smooth Pursuit Movements

Smooth pursuit movements are slow, continuous eye movements used to track moving objects. They help maintain the object's image on the retina, allowing for clear and uninterrupted perception of motion [3].

3.4.4 Blinks

Blinks are brief, typically involuntary closures of the eyelids. Though they can cause data loss in eye-tracking recordings, blinks play a functional role in maintaining eye health and

clarity of vision. Furthermore, blink rate and blink latency are valuable indicators of cognitive load and attentional engagement [3].

3.5 A Historical Perspective on Eye Tracking technologies

Eye tracking has progressed from invasive mechanical methods to non-invasive AI-driven systems, improving accuracy, usability, and applications. Early techniques required physical contact, while modern approaches use infrared cameras and deep learning for real-time gaze tracking. Today, eye tracking is widely used in research, technology, and everyday applications, shaping the future of human-computer interaction, marketing, and healthcare, and this evolution can be categorized into four key generations:

3.5.1 First Generation: Mechanical Eye Tracking (Early 1900s - Mid 20th Century)

Early eye tracking methods used scleral contact lenses with mirrors and electro-oculography (EOG) to measure eye movements. However, these techniques were highly invasive, requiring direct physical contact with the eye, which made them uncomfortable for participants. They also suffered from low accuracy, with frequent measurement errors and significant signal noise affecting the reliability of the data [1].

Figure 3.2: Electro-OculoGraphy-EOG [09]

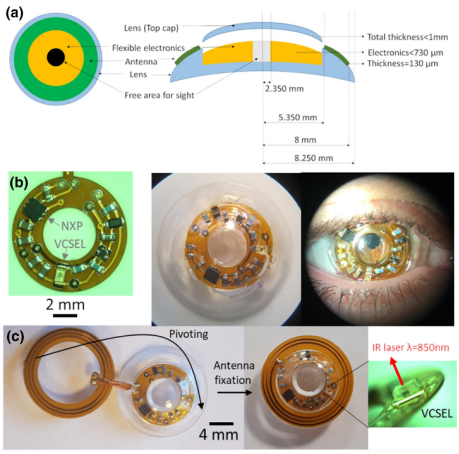


Figure 3.3: Scleral contact lenses [10]

3.5.2 Second Generation: Video-Based Eye Tracking (Late 20th Century)

The second generation of eye tracking introduced the use of infrared (IR) cameras to track pupil and corneal reflections. This advancement made eye tracking non-invasive and much more comfortable for participants compared to earlier methods. It also enabled real-time gaze tracking and led to the development of both head-mounted and remote eye trackers. However, these systems still required extensive calibration and were susceptible to errors caused by head movements, limiting their accuracy in more dynamic settings [1].

Figure 3.4: infrared (IR) cameras [11]

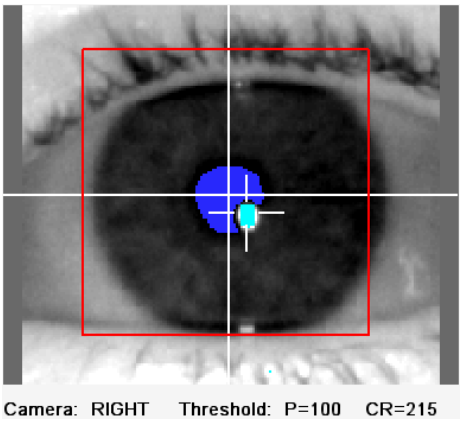


Figure 3.5: (IR) cameras results [12]

3.5.3 Third Generation: Digital and Computational Advancements (2000s - 2010s)

The third generation of eye tracking incorporated machine learning algorithms, high-speed cameras, and improved pupil-tracking software. These advancements led to significantly higher accuracy thanks to better data processing techniques. Portable eye trackers became available, making it possible to conduct studies in real-world settings such as driving and sports. Despite these improvements, the systems were computationally expensive, and some devices remained costly and still required controlled environments to ensure optimal performance [1].

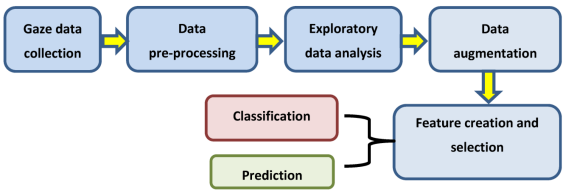


Figure 3.6: Machine learning based analysis [13]

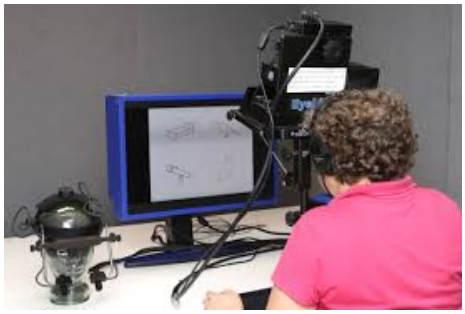


Figure 3.7: High-speed cameras [14]

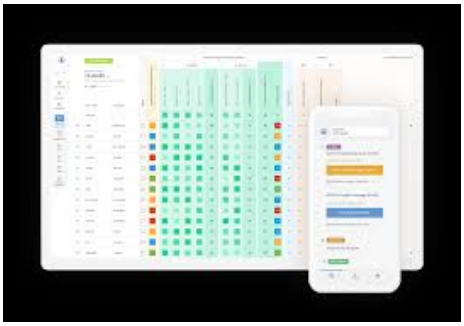
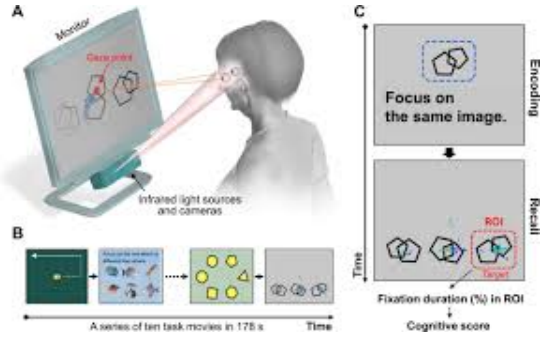


Figure 3.8: pupil tracking software [15]

3.5.4 Fourth Generation: AI Deep Learning (2020s - Present)

The latest generation of eye tracking methods uses AI-driven gaze tracking, wearable eye-tracking glasses, and webcam-based tracking systems. Deep learning models have significantly improved the accuracy of gaze prediction, while webcam-based tracking has made eye tracking more accessible by removing the need for specialized hardware. Additionally, integration with augmented reality (AR), virtual reality (VR), and gaming has opened up new interactive experiences. However, this new wave of technology also faces challenges, including privacy concerns related to AI-based tracking and some accuracy limitations when used in uncontrolled environments [1].



]

Figure 3.9: AI-driven gaze tracking [16]

Figure 3.10: wearable eye-tracking glasses [17]

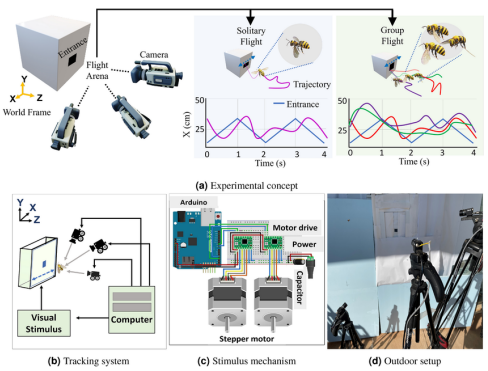


Figure 3.11: Three camera based tracking system [18]

3.6 Key Metrics and Measures

Eye tracking provides several key metrics and measures that are fundamental to understanding visual behavior and cognitive processes during tasks like visual search. they allows researchers to quantify and visualize how individuals interact with visual information,the most common traditional eye tracking metrics are:

3.6.1 Fixations , gaze points and saccades

Gaze points indicate where the eyes are focused at any given moment, with eye trackers often capturing dozens per second depending on the sampling rate, When these points cluster closely in time and space, they form a **fixation**, signaling that the eyes are steadily observing a specific object or area. These fixations are crucial for understanding visual attention. In between fixations are **saccades** quick eye movements that occur, for example, when reading, as the eyes jump from one group of words to another The concept of visual span refers to the amount of surrounding text perceived during a fixation, and it tends to be larger in skilled readers. Conversely, when tracking a moving object like a car, the eyes may engage in smooth pursuit a different type of movement that may still be interrupted by saccades if the object moves unpredictably. Ultimately, analyzing where and how often fixations occur within a scene reveals which areas draw the most attention, laying the foundation for deeper interpretations of user focus and visual behavior [19].

Figure 3.12: gazepath [19]

3.6.2 Areas of Interest (AOI)

Areas of Interest (AOIs) are specific regions within a visual stimulus that researchers define to analyze eye-tracking data more precisely. While not metrics themselves, AOIs serve as boundaries within which various metrics such as time to first fixation, total time spent, number of fixations, and revisit count (can be calculated). For instance, in an image of a person, separate AOIs might be drawn around the face and body to compare attention to each. AOIs are particularly useful for evaluating how different sections of a visual scene, such as parts of a website or elements in an advertisement, capture and hold viewers' attention [19].

Figure 3.13: Magazine AOIs [19]

3.6.3 Time to First Fixation

Time to First Fixation (TTFF) measures how long it takes a viewer to look at a specific Area of Interest (AOI) after a visual stimulus appears. This metric reveals which elements in a scene attract attention first, making it valuable for understanding what draws the eye. TTFF can reflect bottom-up attention, driven by visually striking features like bright colors or movement, or top-down attention, guided by the viewer's goals or expectations. Whether users are naturally drawn to an element or intentionally seek it out, TTFF helps identify which parts of a design or image are prioritized early in the viewing process [19].

3.6.4 First Fixation Duration

First fixation duration refers to the amount of time a participant's eyes remain focused on a specific area during their very first glance. It typically follows a saccade and can offer valuable insights when combined with the Time to First Fixation (TTFF). For example, a short TTFF followed by a long first fixation suggests that a region captured attention quickly and strongly. Measured within defined Areas of Interest (AOIs), this metric helps identify which elements in a visual scene leave a strong first impression, making it especially useful for comparing the initial impact of different regions [19].

3.6.5 Time spent (Dwell Time)

Dwell Time, or time spent, refers to the total duration a viewer's gaze remains on a specific Area of Interest (AOI). This measure is useful for evaluating engagement or focus on particular parts of a visual scene. Longer dwell times often suggest greater interest or intentional attention (top-down), while shorter times may indicate lower relevance or competing distractions elsewhere [19].

3.6.6 Revisits

The number of revisits refers to how often a participant returns their gaze to a specific Area of Interest (AOI) after looking elsewhere. This metric reveals which parts of a visual scene repeatedly attract attention, whether due to interest, confusion, or other cognitive responses. While eye tracking alone can't determine the emotional cause behind a revisit, it highlights ar-

eas that merit further investigation through complementary methods, helping researchers identify elements that consistently capture or retain visual attention [19].

3.6.7 Heatmaps

Heatmaps are a popular way to visualize where people focus their attention during eye tracking studies. By overlaying a color gradient typically using red for the most viewed areas, followed by yellow and green on an image or stimulus, heatmaps provide an immediate, intuitive picture of how gaze points are distributed. This makes it easy to identify which elements of a scene attracted the most attention. Heatmaps can be generated for individuals or groups, allowing researchers to compare how different users interact with the same visual content [19].

Figure 3.14: Car heatmap [19]

3.7 Related work

3.7.1 Classic eye tracking

The systematic study of eye movements dates back to the early 20th century, with foundational contributions from pioneers such as Edmund B. Huey, Raymond Dodge, and T.S. Cline. Huey (1908) was among the first to rigorously examine reading behavior by tracking eye movements [20], while Dodge and Cline (1901) developed photographic techniques to measure the angular velocity of ocular motion [21]. These early studies laid the groundwork for later methodological advancements. Researchers like G.T. Buswell (1935) extended this work

by investigating how individuals visually explore complex stimuli such as pictures, providing early insights into visual attention patterns [22].

Over the years, various eye-tracking methodologies have evolved. Early technologies included Electro-Oculography (EOG), which detects corneo-retinal potential differences; the Scleral Search Coil method, which uses a contact lens embedded with a wire coil to track eye position in a magnetic field; and Photo-Oculography (POG) or Video-Oculography (VOG), which utilize light reflections from the eye surface to record movement (Young Sheena, 1975). These technologies enabled researchers to quantify eye behaviors with increasing precision [23].

A landmark in the cognitive interpretation of eye movements was established by A.L. Yarbus (1967), who demonstrated that eye movements are influenced not merely by visual stimuli, but also by the viewer's cognitive goals and task demands. This insight laid the foundation for understanding scanpaths as task-driven, rather than purely stimulus-driven [24]. Building on this, Just and Carpenter (1980) introduced the eye-mind hypothesis, suggesting a tight link between eye fixations and real-time cognitive processing during tasks such as reading and problem solving [25].

Kenneth Rayner (1998) later synthesized decades of research, offering a comprehensive review of eye movement behavior in reading and information processing. His work highlighted the centrality of fixations, saccades, and regressions in understanding visual cognition [26]. Early eye-tracking studies often focused on quantifying basic metrics such as fixation count, mean fixation duration, saccade amplitude, and scanpath sequence, as well as how attention is distributed across predefined Areas of Interest (AOIs). These traditional metrics remain fundamental in eye-tracking research today.

Early eye-tracking systems, such as those developed by ISCAN, were hardware-intensive and required manual calibration and serial communication for data transfer. Despite these limitations, they enabled researchers to explore foundational questions about oculomotor control, perception, and attention. Together, these early technologies and theoretical insights laid the groundwork for the sophisticated, real-time eye-tracking systems used in modern cognitive, psychological, and human-computer interaction research [27].

3.7.2 Modern Eye Tracking

Today, eye-tracking technology has advanced significantly, becoming a widely used tool in diverse research domains including psychology, neuroscience, education, marketing, and human-computer interaction, and it has evolved significantly, moving from hardware-dependent solutions to AI-driven, camera-based methods

AI in Retinal Disease Screening: Artificial intelligence (AI) has become a valuable tool for screening retinal diseases such as diabetic retinopathy, age-related macular degeneration (AMD), retinopathy of prematurity, and glaucoma. Deep learning (DL) techniques, particularly convolutional neural networks (CNNs), have shown high accuracy in diagnosing these conditions. For example, DL models have been developed to predict the need for anti–vascular endothelial growth factor (anti-VEGF) therapy in AMD patients by analyzing retinal imaging data with impressive predictive accuracy [28].

Deep Learning for Glaucoma Detection: There is an increasing emphasis on using deep learning to automate glaucoma detection. Researchers have designed CNN-based architectures capable of classifying glaucoma from optical coherence tomography (OCT) images . These models can also highlight regions of interest (ROIs) in OCT scans that are most relevant for detecting structural changes associated with glaucoma, supporting clinical decision-making and longitudinal monitoring [29].

Attention-Guided CNNs: Li et al. (2019) proposed an attention-guided convolutional neural network trained using simulated clinician eye-tracking data to identify diagnostically relevant regions in fundus images. Their model achieved a glaucoma detection accuracy of over (95) , emphasizing the value of incorporating human visual attention into the training process to guide DL models [30].

LEyes: A Lightweight Framework for Deep Learning-Based Eye Tracking Using Synthetic Eye Images

This study introduces LEyes, a framework designed to address the scarcity of diverse training data in eye-tracking research. By generating synthetic eye images that model essential features for gaze estimation, LEyes enables the training of deep learning models that generalize well across different datasets. The framework demonstrates competitive performance in

pupil and corneal reflection localization, even outperforming some industry-standard eye trackers using more affordable hardware [31].

3.8 Eye tracking databases

Eye-tracking databases provide valuable resources for studying visual attention, gaze behavior, and cognitive processing across various fields. These datasets support research in human-computer interaction, psychology, AI, and visual search, helping improve usability, machine learning models, and user experience design. Some well-known eye-tracking datasets include :

3.8.1 General Eye-Tracking Databases:

GazeBase is a comprehensive dataset used extensively in psychology and neuroscience to study eye movement behavior under various experimental conditions. It captures a wide range of gaze patterns across different tasks and user states, enabling researchers to analyze attention, fatigue, and cognitive load in naturalistic settings [32].

Tobii Pro Research Data is provided by Tobii, a leading manufacturer of eye-tracking hardware. This dataset is widely used in human-computer interaction (HCI) and user experience (UX) research. It includes data from usability testing, reading behavior studies, and cognitive workload assessments. Its high-quality recordings from Tobii's proprietary devices make it particularly valuable for industrial and academic research [33].

MPIIGaze is a benchmark dataset in the field of machine learning and artificial intelligence. Designed for appearance-based gaze estimation in real-world settings, it features head pose and gaze direction data collected from users in everyday environments. MPIIGaze has been instrumental in training and evaluating deep learning models for gaze prediction under varying lighting and head movement conditions [34].

EYEDIAP serves the computer vision and augmented reality (AR) research communities. It includes gaze data captured with RGB and depth cameras, making it ideal for 3D gaze estimation tasks. The dataset supports the development of robust gaze tracking systems in spatial computing and immersive applications [35].

OpenGaze is an open-source dataset aimed at facilitating research in assistive technology and HCI. It includes recordings for evaluating gaze-based user interfaces, with a focus on accessibility and natural interaction design. This dataset supports studies on gaze-based control systems for users with motor impairments or alternative input needs (OpenGaze Dataset) [36].

Table 3.1: General Eye-Tracking Databases Comparison

Dataset	**Primary Use Case**	**Modality**	**Tasks Included**	**Participants**	**Environment**	**Open Access**
Gaze Base[32]	Cognitive & behavioral study	Eye-tracker (binocular)	Fixation, pursuit, reading, etc.	∼322	Controlled + Naturalistic	Yes
Tobii Pro[33]	HCI, UX, reading analysis	Tobii hardware (proprietary)	Reading, usability, cognitive tasks	Varies	Lab, commercial	No
MPII Gaze[34]	Gaze estimation (ML/AI)	RGB webcam	Free-viewing, head pose variation	15	In-the-wild (home)	Yes
EYE DIAP [35]	3D gaze & pose estimation	RGB-D (depth + color)	Gaze & head pose, speaking tasks	16	Lab (controlled)	Yes
Open Gaze [36]	Accessibility, Assistive Tech	Eye-tracker + webcam	UI interaction, gaze control	Small group	Controlled	Yes

3.8.2 Eye-Tracking Databases for Visual Search Tasks

These datasets focus on how users search for objects or patterns in visual scenes, widely used in cognitive science, AI, and UX design.

COCO-Search18 is a large-scale dataset designed to facilitate the study of goal-directed visual search in naturalistic scenes. Based on the MS COCO image collection, it includes over 6,000 images and records human fixation data during search tasks for everyday objects. The

dataset supports computational modeling of human attention and behavior during search in realistic scenarios [37].

OSIE (Object and Semantic Images Eye-tracking dataset) focuses on the influence of object complexity and semantics on visual attention. It tracks human gaze behavior across images with varying levels of object density and meaning, providing valuable insight into how cognitive and visual factors drive fixation patterns [38].

The MIT Saliency Benchmark, often used in conjunction with the MPIIGaze dataset, serves as a standard resource for evaluating gaze prediction and saliency models in real-world environments. These datasets are widely applied in machine learning and deep learning-based gaze estimation research, helping benchmark algorithms for performance in diverse visual contexts [39].

FIFA (Feature Integration for Fixation Analysis) investigates how humans perform visual search and fixation tasks in controlled, experimental environments. The dataset is designed to explore theories such as Feature Integration Theory, enabling deeper understanding of attention mechanisms and fixation strategies [40].

The Visual Search in Cluttered Scenes (VSCS) dataset is aimed at examining how people locate targets in highly cluttered visual environments. It captures the dynamics of visual search in settings where distractors are numerous and similar to the target, which is essential for modeling real-world attention challenges [41].

UEyes dataset is a large-scale eye-tracking dataset comprising 1,980 UI screenshots across four categories—Webpage, Desktop UI, Mobile UI, and Poster—collected from 62 participants in a controlled lab setting. It includes gaze data, fixation logs, and metadata, enabling research into user attention and visual search behavior across diverse interface types [42].

Table 3.2: Eye-Tracking Databases for Visual Search Tasks Comparison

Dataset	**Primary Focus**	**Scene Type**	**Data Type**	**Participants**	**Search Task**	**Open Access**
COCO-Search18 [37]	Goal-directed search in scenes	Natural images (MS COCO)	Fixations, bounding boxes	∼10	Search for specific objects	Yes
OSIE [38]	Visual attention & semantics	Naturalistic object scenes	Gaze, object masks, semantics	15	Free viewing with semantic complexity	Yes
MIT Saliency Benchmark [39]	Saliency & attention modeling	Diverse image categories	Fixation maps, saliency maps	∼15–20	Free viewing of diverse stimuli	Yes
FIFA [40]	Visual feature integration	Synthetic scenes	Fixation locations, response times	∼30	Search based on visual features	Yes
VSCS [41]	Search in cluttered environments	Synthetic + natural clutter	Fixations, search performance	∼20	Search with many distractors	Yes
UEyes[42]	Visual search in UI contexts	Web, Desktop, Mobile, Posters	Fixations, scanpaths, metadata	62	Search tasks across multiple UI categories	Yes

3.9 Conclusion

In this chapter, we explored the core concepts and technologies that form the basis of eye tracking research. From understanding the fundamental types of eye movements to reviewing

historical milestones and technological advancements, we gained insight into how eye tracking systems capture and interpret human gaze behavior. Key metrics and measures were highlighted as essential tools for analyzing visual attention. Additionally, a review of related work and important eye tracking databases emphasized the growing body of knowledge in this area. This foundation prepares readers for deeper exploration into the applications and future directions of eye tracking in research and industry.

Chapter 4

Methodology

4.1 Introduction

In this chapter, we present the methodology and design of our study on eye tracking in simple visual search tasks. We begin by describing the dataset and Deep learning architecture with metrics that we used than we talked about the proposed eye tracking system , highlighting the essential steps required to analyze eye movement data effectively. These steps include preprocessing and cleaning of raw gaze data, followed by model training and feature extraction, which allows us to capture relevant visual attention patterns. Finally, we conduct data analysis and visualization to interpret the results and gain insights into user behavior during visual search. The ultimate goal is to develop a reliable system for studying and understanding visual attention in controlled search scenarios.

4.2 Eye tracking dataset

- **THE DATASET:UEYES**

The UEyes dataset consists of 1,980 UI screenshots along with corresponding metadata and eye-tracking logs from 62 participants, gathered in a controlled laboratory environment using an advanced eye tracker. The dataset includes 495 screenshots for each of the following UI types: Webpage, Desktop UI, Mobile UI, and Poster. For the Webpage category, 494 screenshots were sourced from the Alexa 500 dataset, 1,507 images from the Visual Complexity and

Aesthetics dataset, and 200 images from the Imp1k dataset, with an additional 103 webpage screenshots captured to expand the image set. The Desktop UI category includes 51 images from the Waltteri Github desktop UI dataset, supplemented by 303 desktop UI screenshots collected under specific criteria to ensure diversity. The Mobile UI category comprises 1,761 images selected from a pool of 46,064 images in the RICO dataset, with 42 additional mobile UI images representing diverse app categories such as school, library, music, and settings. Lastly, the Poster category features 200 ad images and 198 infographics from the Imp1k dataset, along with 103 additional posters chosen for their distinctiveness or widespread relevance. These extra images were carefully selected to ensure diversity and representational breadth, balancing the dataset with images that reflect common, everyday designs. To ensure consistency, any images containing inappropriate content, such as pornography, were removed. The remaining images were grouped into 55 "image blocks" for participant assessment, each containing nine images from each UI type, totaling 36 images per block. During data collection, the screen angle was adjusted for each participant to match their typical viewing conditions, with participants seated at a distance of 50–65 cm from the screen. This uniform setup ensured that the comparison across different UI types was fair, minimizing the impact of eye tracker technology's accuracy on the results, particularly for mobile UIs [42].

Table 4.1: Summary of the UEyes Dataset Content and Characteristics

Attribute	**Description**
Purpose	Eye-tracking data for understanding visual saliency across different UI types
Number of Participants	62 participants
Number of UI Screenshots	1,980 UI screenshots in total
UI Types Covered	4 types: Webpage, Desktop UI, Mobile UI, and Poster
Screenshots per UI Type	495 screenshots for each UI type
Data Collected	- High-fidelity in-lab eye-tracking data (gaze fixations) - Multi-duration saliency maps - Scanpaths (fixation sequence and timing)
Data Format	Logs containing fixation locations, durations, and temporal order
Biases Analyzed	Location bias (center/horizontal), color bias, saccade angles, revisit patterns
Applications	- Training/evaluation of saliency models - UI design improvement - Understanding gaze differences across UI types
Dataset Availability	Publicly available with full metadata and eye-tracking logs
Significance	One of the largest multi-UI-type eye-tracking datasets, supports generalizable saliency modeling

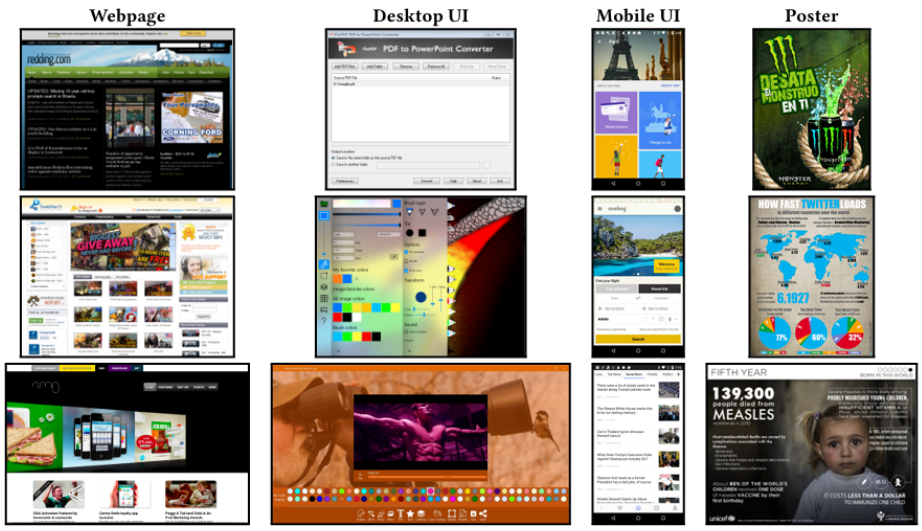


Figure 4.1: Examples of user interfaces in the UEyes dataset [42]

The UEyes dataset was chosen for this study because it offers a large, diverse collection of real-world UI screenshots across multiple interface types—web, desktop, mobile, and posters—making it ideal for analyzing visual search behavior. Its high-quality eye-tracking data, collected in a controlled environment, ensures consistency and reliability. The dataset's variety and structured design allow for fair comparison across UI types and support the study of how users visually explore simple yet realistic visual scenes.

4.3 Deep learning architecture

Deep learning is an advanced subset of machine learning that employs neural networks with multiple layers to capture and learn intricate patterns from large volumes of data. In contrast to traditional machine learning techniques, which often depend on manually engineered features, deep learning models can automatically extract meaningful features directly from raw input through successive layers of nonlinear transformations.

This capability makes deep learning highly effective in fields such as image analysis, speech recognition, and natural language processing. The foundation of this approach lies in artificial neural networks, which are inspired by the structure and functioning of the human brain—comprising layers of interconnected neurons that simulate biological information processing. The term "deep" signifies the presence of many such layers, enabling the model to learn and represent data in a hierarchical manner [43].

Deep learning architectures consist of layered neural networks that learn hierarchical representations of data. They typically include an input layer, several hidden layers for feature extraction and abstraction, and an output layer for prediction. These hidden layers apply operations like weighted summations, nonlinear activations (like ReLU, Sigmoid, Tanh), and regularization techniques such as normalization and dropout to improve model performance. Depending on the data type and task, various architectures are employed for example, Feedforward Neural Networks (FNNs) for structured data, Recurrent Neural Networks (RNNs) for sequential data, and Transformers for long-range dependencies. Among these, Convolutional Neural Networks (CNNs) are especially prominent in visual data processing. CNNs use convolutional layers with learnable filters to extract spatial features, pooling layers to reduce dimensionality, and fully connected layers for final decision-making.

This architecture enables the model to learn both low-level and high-level patterns in images while maintaining computational efficiency through weight sharing. As a result, CNNs are widely used in applications such as image classification, object detection, and computer vision tasks [44].

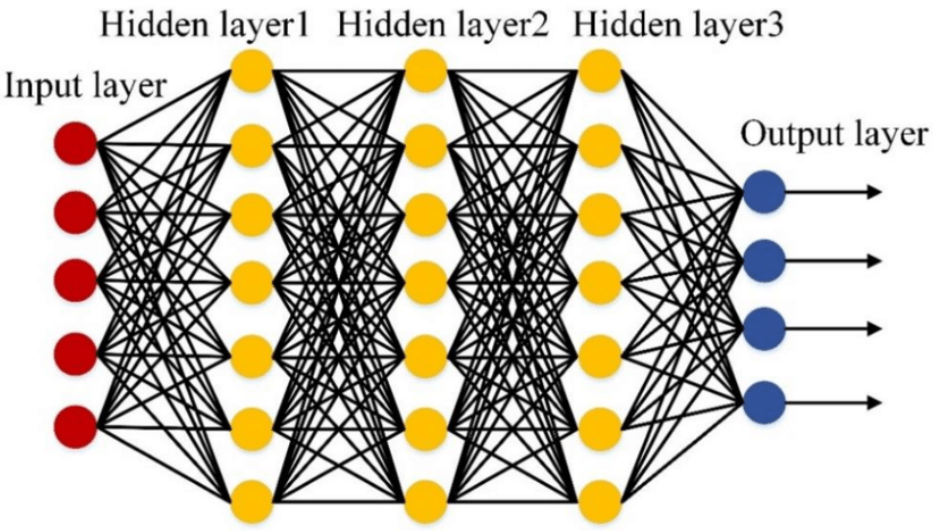


Figure 4.2: General Deep learning architecture [45]

One of the most influential CNN-based architectures is the U-Net model.

4.3.1 U-Net model

The U-Net is a convolutional neural network (CNN) architecture specifically designed for image segmentation, distinguished by its U-shaped structure that comprises a contracting path (encoder) and an expansive path (decoder). The contracting path captures increasingly abstract features by repeatedly applying 3×3 convolutions, ReLU activations, and 2×2 max pooling operations, which reduce spatial resolution while doubling the number of feature channels. In contrast, the expansive path restores spatial resolution through transposed convolutions that halve the number of channels and incorporate skip connections from the encoder. These skip connections concatenate high-resolution features from earlier layers with the upsampled output, enabling the model to recover fine spatial details lost during downsampling. The network concludes with a 1×1 convolution that maps features to the required number of segmentation classes. Unlike traditional CNNs, U-Net omits fully connected layers, relying solely on convolutional operations to perform precise, pixel-level predictions efficiently. Its ability to combine deep feature extraction with spatial detail recovery makes it especially effective for tasks such as biomedical image segmentation, where accurate localization is essential [46].

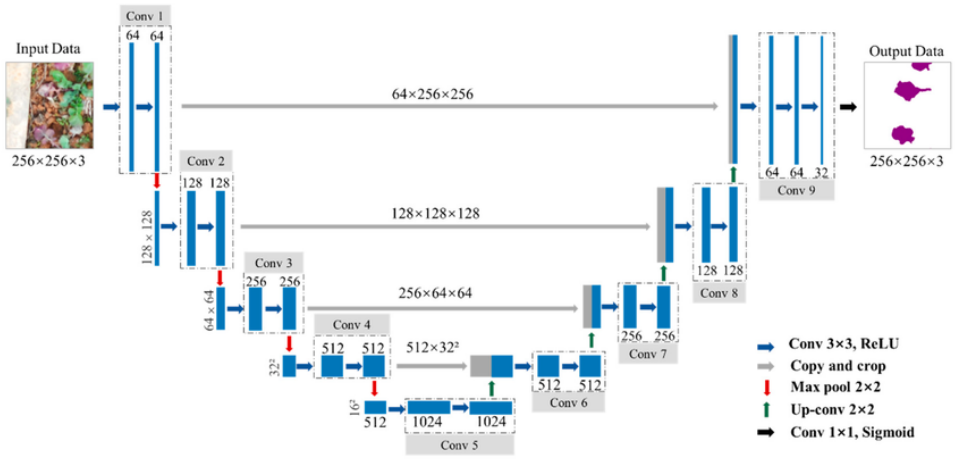


Figure 4.3: U-NET architecture for eye tracking system [47]

4.4 Quantitative Evaluation of Visual Saliency Maps Using Eye-Tracking Data

In deep learning, evaluating model performance relies on a variety of metrics tailored to different tasks such as regression, classification, and saliency prediction..

4.4.1 Regression Metrics (Pixel-based error)

Mean Absolute Error (MAE):

MAE calculates the average absolute difference between predicted values $\hat{y}_i$ and true values y_i :

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{4.1}$$

Mean Squared Error (MSE):

MSE measures the average squared difference, placing greater emphasis on larger errors:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{4.2}$$

These metrics provide intuitive error quantifications for continuous prediction tasks, with MSE being more sensitive to outliers [48].

4.4.2 Probabilistic / Distribution Metrics

Kullback-Leibler Divergence (KLD): In tasks involving probabilistic outputs, such as saliency map prediction, the Kullback-Leibler Divergence (KLD) quantifies the difference between predicted (P) and ground truth (Q) distributions:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{4.3}$$

Lower KLD values indicate better alignment between distributions [49].

Similarity (SIM): The Similarity (SIM) metric measures the overlap between two normalized saliency maps P and Q by summing the minimum values pixel-wise:

$$\text{SIM} = \sum_i \min(P(i), Q(i)) \tag{4.4}$$

Ranging from 0 (no overlap) to 1 (perfect match)

4.4.3 Fixation Prediction Metrics

Correlation Coefficient (CC):

The Correlation Coefficient (CC) assesses the linear relationship between predicted and true saliency maps by measuring covariance normalized by their standard deviations:

$$\text{CC} = \frac{\sum_i (\hat{S}_i - \mu_{\hat{S}})(S_i - \mu_S)}{\sqrt{\sum_i (\hat{S}_i - \mu_{\hat{S}})^2 \cdot \sqrt{\sum_i (S_i - \mu_S)^2}}} \tag{4.5}$$

Where μ denotes the mean values of the maps

4.4.4 General Classification

Accuracy In classification tasks, Accuracy remains a fundamental measure representing the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i) \tag{4.6}$$

Where 1 is an indicator function for correct predictions [50].

Together, these metrics provide comprehensive tools to assess the quality and reliability of deep learning models across diverse applications.

4.5 Proposed Eye tracking system

The proposed eye-tracking system leverages deep learning techniques to predict saliency maps from visual stimuli using the uEyes dataset. This system employs a U-Net architecture, a type of convolutional neural network designed for pixel-wise predictions, making it ideal

for saliency prediction tasks. The uEyes dataset, which consists of images from various categories such as desktop, mobile, poster, and web, includes corresponding eye-tracking data in the form of saliency maps generated at different durations (1s, 3s, and 7s). These maps represent areas of high attention during visual searches, providing valuable insights into human visual behavior. The model is trained on preprocessed images resized to 256x256 pixels, with normalization applied to both images and their respective saliency maps. Custom evaluation metrics such as Mean Absolute Error(MAE),Mean Squared Error(MSE),Kullback-Leibler Divergence (KLD), Pearson's Correlation Coefficient (CC), and Histogram Similarity (SIM) are used to assess the accuracy of the model's predictions in comparison to the ground truth saliency maps. By employing a U-Net architecture, which incorporates an encoder-decoder structure with skip connections, the model efficiently captures both fine details and contextual information in the visual inputs, improving the accuracy of saliency prediction. The use of the uEyes dataset, combined with advanced deep learning models, enables the development of an effective and robust eye-tracking system for understanding human visual attention in various real-world scenarios.

4.5.1 Preprocessing and Cleaning techniques

The preprocessing and cleaning techniques are essential to ensure high-quality input data for the model. In our system Before actual preprocessing begins, several preparatory steps are performed to ensure that the data is complete and well-organized. First, metadata is loaded, which includes the names and types of the images in the dataset. Then, the system filters out only valid image–saliency map pairs by verifying that both the image and its corresponding saliency map exist. This step effectively handles any missing or incomplete data. Once valid pairs are identified, the dataset is divided into training (80%) and validation (20%) subsets to allow for reliable evaluation of the model's performance.

After these preparatory steps, the core preprocessing phase is applied to each image. All images are resized to a consistent resolution of 256×256 pixels to ensure uniformity in input dimensions. Pixel values are then normalized to a [0, 1] range, which improves the model's ability to learn efficiently by standardizing input values. Additionally, images are converted from BGR to RGB format to maintain color accuracy, as required by most deep learning models

Finally, a custom data generator is used to load the images and their corresponding saliency

maps in batches during training. This not only conserves memory but also enables efficient data feeding to the model. Through this structured preprocessing pipeline, the system ensures optimal conditions for model training while maintaining data quality and efficiency.

Figure 4.4: Preprocessed image with their target silency map

4.5.2 U-Net model training

4.5.2.1 Inputs images :

The model begins with an input layer that accepts RGB images of size 256×256 pixels. This standardized input shape allows the model to process and learn from images in a consistent spatial format.

4.5.2.2 Encoder part

The encoder path, also known as the contracting path. This part of the network is composed of multiple convolutional blocks(blocks 1 to 4), where all the Feature extraction happens, the model extracts low-level features such as edges, corners, and textures using small receptive fields. As the image passes through deeper layers, these features are transformed into high-level representations, capturing shapes, regions, and semantic context related to visual attention. Each encoder block uses two convolutional layers with ReLU activations to emphasize non-linear patterns, followed by max pooling to reduce spatial dimensions and focus on the most

prominent structures. Importantly, the number of feature channels doubles at each depth level (e.g., 64, 128, 256, 512), allowing the model to learn a wide range of visual attributes. These extracted features are essential for tasks like saliency prediction, where understanding both fine-grained details and overall scene context is necessary to localize areas of interest effectively.

4.5.2.3 Middle / Bottleneck

After the encoder, the U-Net reaches the "middle" bottleneck (block 5), where two convolutional layers with 1024 filters are applied, followed by dropout to reduce overfitting. This layer represents the deepest feature extraction in the network, with the smallest spatial resolution but the highest number of feature channels, capturing the most complex and abstracted information from the input in this way :

4.5.2.4 Decoder (Upsampling Path)

The decoder of U-Net (blocks 6 to 9) mirrors the encoder and focuses on restoring the image's spatial resolution while combining it with high-level features via skip connections. Each decoder block begins with upsampling through transposed convolutions, followed by feature concatenation from the corresponding encoder layer. This fusion of low-level spatial detail and high-level context helps the model generate accurate, localized output maps.

4.5.2.5 Output Layer

The final layer applies a 1×1 convolution with a sigmoid activation to produce a single-channel output (a saliency map) where each pixel value indicates the predicted attention intensity. The sigmoid ensures outputs fall between 0 and 1, making them interpretable as probabilities.

Figure 4.5: Extract from the model summary

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 256, 256, 3)	0	-
conv2d (Conv2D)	(None, 256, 256, 64)	1,792	input_layer[0][0]
conv2d_1 (Conv2D)	(None, 256, 256, 64)	36,928	conv2d[0][0]
max_pooling2d (MaxPooling2D)	(None, 128, 128, 64)	0	conv2d_1[0][0]
conv2d_2 (Conv2D)	(None, 128, 128, 128)	73,856	max_pooling2d[0]…
conv2d_3 (Conv2D)	(None, 128, 128, 128)	147,584	conv2d_2[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 64, 64, 128)	0	conv2d_3[0][0]
conv2d_4 (Conv2D)	(None, 64, 64, 256)	295,168	max_pooling2d_1[…
conv2d_5 (Conv2D)	(None, 64, 64, 256)	590,080	conv2d_4[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 32, 32, 256)	0	conv2d_5[0][0]
conv2d_6 (Conv2D)	(None, 32, 32, 512)	1,180,160	max_pooling2d_2[…
conv2d_7 (Conv2D)	(None, 32, 32, 512)	2,359,808	conv2d_6[0][0]
dropout (Dropout)	(None, 32, 32, 512)	0	conv2d_7[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 16, 16, 512)	0	dropout[0][0]
conv2d_8 (Conv2D)	(None, 16, 16, 1024)	4,719,616	max_pooling2d_3[…
conv2d_9 (Conv2D)	(None, 16, 16, 1024)	9,438,208	conv2d_8[0][0]
dropout_1 (Dropout)	(None, 16, 16, 1024)	0	conv2d_9[0][0]
conv2d_transpose (Conv2DTranspose)	(None, 32, 32, 512)	2,097,664	dropout_1[0][0]

4.5.3 Data Analysis and Visualization

After training the U-Net model for saliency prediction, the evaluation phase rigorously assesses performance on a separate test set by computing a comprehensive suite of metrics including MAE, MSE, accuracy, and specialized saliency metrics like KLD, CC, and SIM, which quantify how well the predicted attention maps match human gaze patterns. Input images and their ground truth heatmaps are preprocessed, normalized, and resized to model dimensions before prediction. Visualization plays a key role in both monitoring training and interpreting results, training and validation metric trends are plotted over epochs to reveal learning progress and potential overfitting, while side-by-side comparisons of original images, ground truth saliency maps, and model predictions using heatmap color coding provide qualitative insight into model accuracy. This entire pipeline from data preprocessing through training, evaluation, metric calculation, and visualization is designed for systematic, transparent model development and iterative improvement, with all results saved for documentation and future

analysis.

4.6 Conclusion

This chapter detailed the step-by-step approach used to build and evaluate the proposed eye-tracking system. From data preparation to model training and analysis, each component was carefully designed to capture and interpret user gaze behavior. The combination of deep learning techniques and well-defined metrics allowed for accurate saliency prediction and meaningful insights into visual attention during search tasks.

Chapter 5

Experimental Results

5.1 Introduction

This chapter provide the technical details of the experiment and present the practical outcomes of the proposed eye-tracking system for visual search analysis using deep learning.

5.2 Technical Requirements

The technical requirements for this project are:

5.2.1 Hardware Requirements

The hardware environment in which our application was developed is characterized by:

Table 5.1: System Hardware Specifications

Device 1	
Device Name	DESKTOP-DETNVLO
Processor	Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz 1.90GHz x64
Installed RAM	8.00 GB
Operating System	64-bit
Device 2	
Device Name	DESKTOP-7CLFBIH
Processor	Intel(R) Core(TM) i5-5300U CPU @ 2.30GHz
Installed RAM	8.00 GB (7.88 GB usable)
Operating System	64-bit, x64-based p

5.2.2 Software Requirements

- **Programming Language:**

To implement our eye-tracking system,we chose to use Python (3.7.3),Python is a high-level, object-oriented programming language known for its simplicity and ease of use. Its interactive and interpreted nature makes it an ideal choice, as instructions are translated into machine code that the computer can execute in real time. Python is also renowned for being easier to learn compared to other languages, enabling users to develop high-quality programs efficiently.

- **Code editor:**

To edit the code for our system, we used **Google Colab** which is a cloud-based platform that allows users to write and run Python code in Jupyter notebooks, with free access to GPUs

- **Raport editor:**

To write the raport of our project system we used **Overleaf** A cloud-based collaborative platform for writing, editing, and publishing LaTeX documents.

- **Libraries and packages:**

In this project we used a combination of powerful Python libraries and built-in modules to handle data processing, image analysis, machine learning, and visualization

Matplotlib: Library for creating static and interactive plots and graphs.

Opencv-python (cv2): Tools for image processing and computer vision tasks.

Scikit-image: Collection of image processing algorithms for filtering and segmentation.

Pandas: Provides data structures and functions for easy data manipulation and analysis.

Tensorflow: Framework for building and training machine learning and deep learning models.

Seaborn: Statistical data visualization library built on matplotlib for attractive charts.

Albumentations: Fast image augmentation library to enhance training datasets.

Numpy: Core package for numerical computations with support for arrays and matrices.

Sklearn model selection: Tools for splitting datasets and evaluating models.

Tqdm: Adds progress bars to loops for monitoring execution.

Os: Built-in Python module for interacting with the operating system.

Glob: Built-in module to find files using pattern matching.

Random: Built-in module for generating random numbers and selections.

Google.colab.drive: Module for accessing Google Drive files within Colab notebooks.

Gradio: A Python library that allows to quickly create interactive web interfaces for machine learning models and data science projects.

5.3 Application GUI

The graphical user interface (GUI) of our eye-tracking system is developed to provide a user-friendly and intuitive experience for visual attention prediction.

When users open the application, they are greeted by a Home page that highlights the system's main feature:

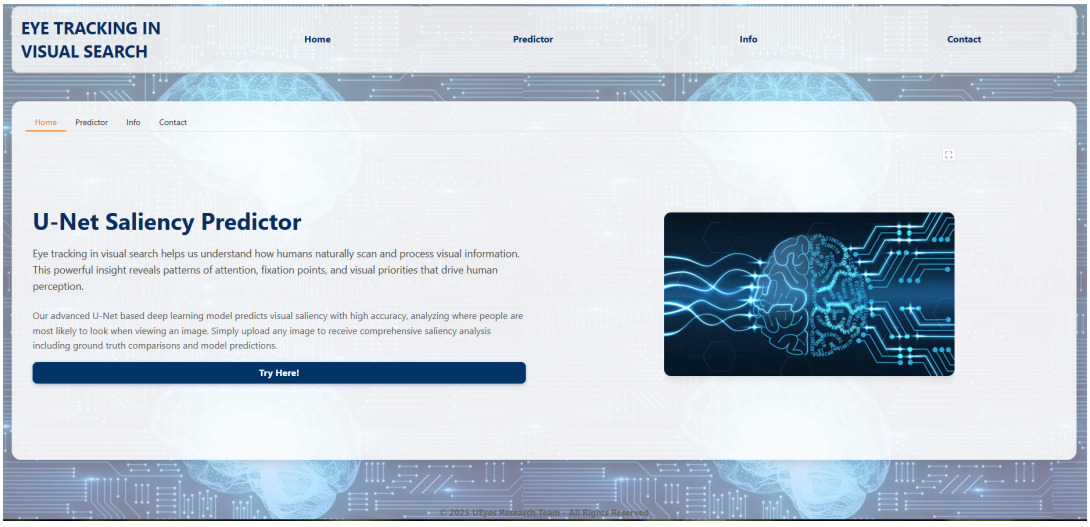


Figure 5.1: Home page

By navigating to the "Predictor" tab, users access the core functionality of the system. They can upload an image by dragging and dropping or selecting a file manually:

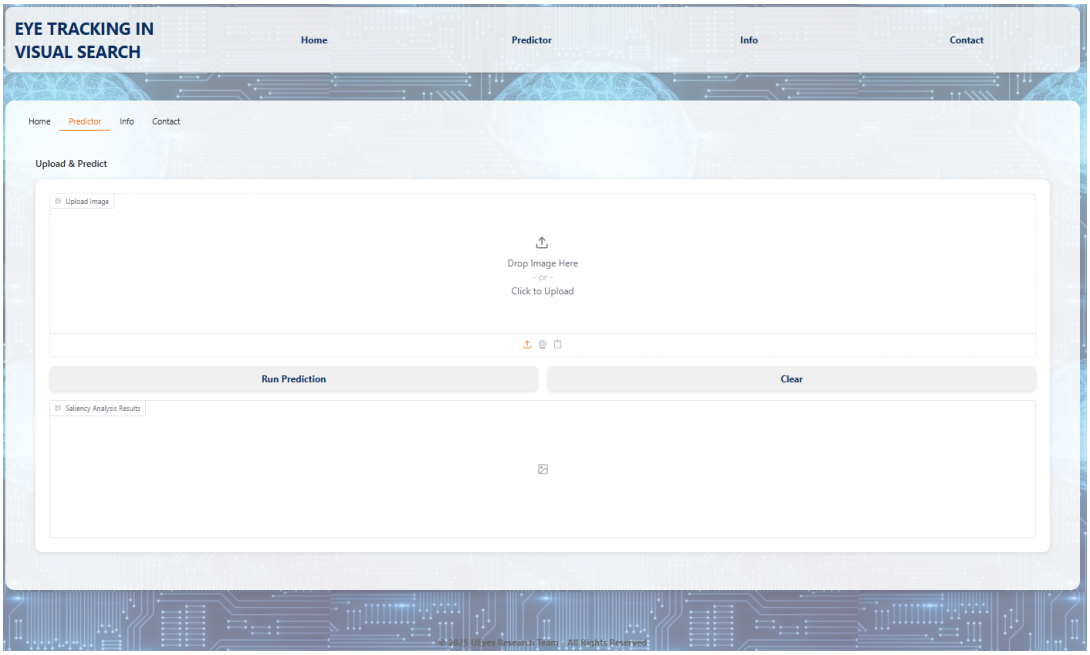


Figure 5.2: Predictor Tab

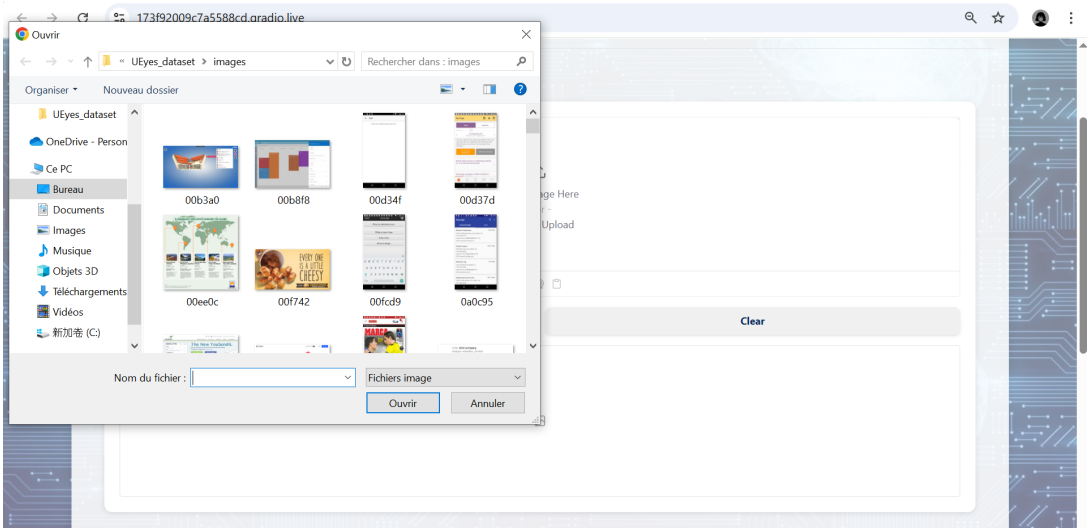


Figure 5.3: Uploading images

Once uploaded, the image is displayed in the interface, and users can initiate the prediction process by clicking the "Run Prediction" button:

Figure 5.4: Displayed image

The system then generates and displays the saliency results,ground truth, prediction, and overlay side by side for intuitive comparison:

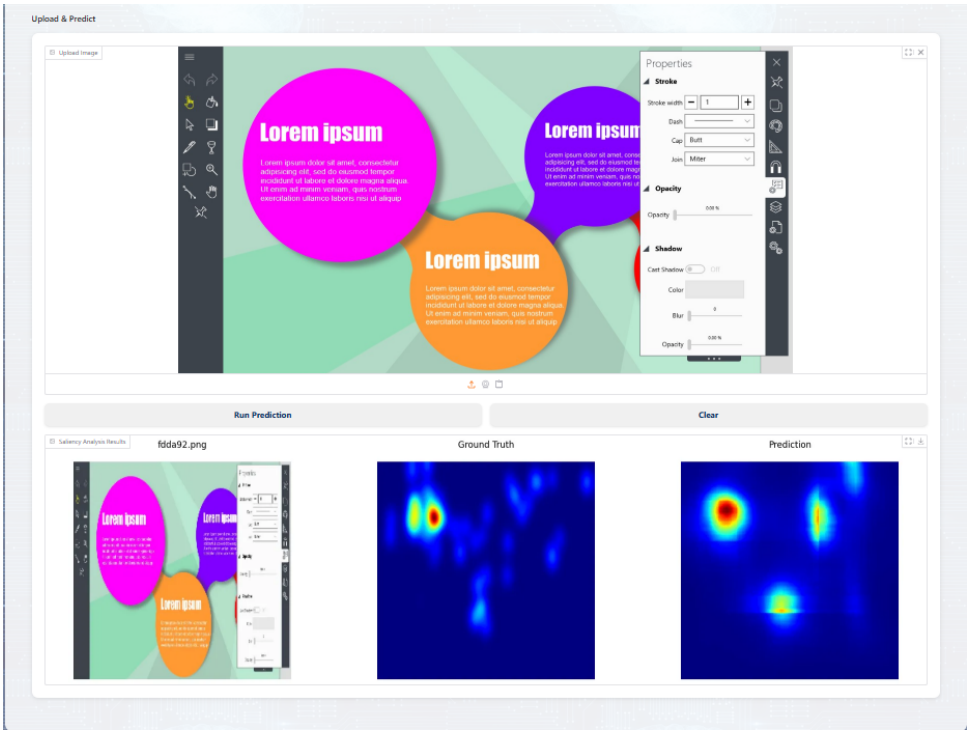


Figure 5.5: The saliency results

Users can also reset the interface with the clear botton and begin a new prediction cycle

with ease.

The "Info" tab offers comprehensive details about the model powering the system. It explains that the solution is based on a U-Net convolutional neural network (CNN) with an encoder-decoder structure, and evaluates performance using metrics such as Accuracy, CC, KLD, NSS, and SIM. It also notes that the model is trained on the UEyes dataset, built using TensorFlow, and the interface is implemented via Gradio. This section further credits the contributors behind the project.

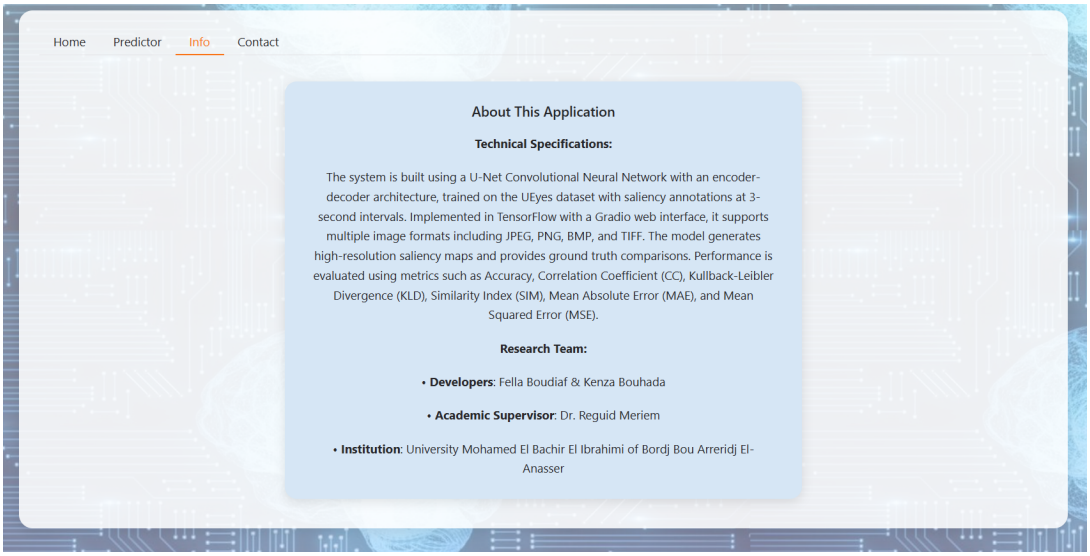


Figure 5.6: Info Tab

The Contact Interface provides a clear and minimal section containing essential contact information such as email addresses, social media handles, a phone number, and a GitHub link. Designed for simplicity, it ensures easy access and navigation for users seeking to get in touch or return to other parts of the application.

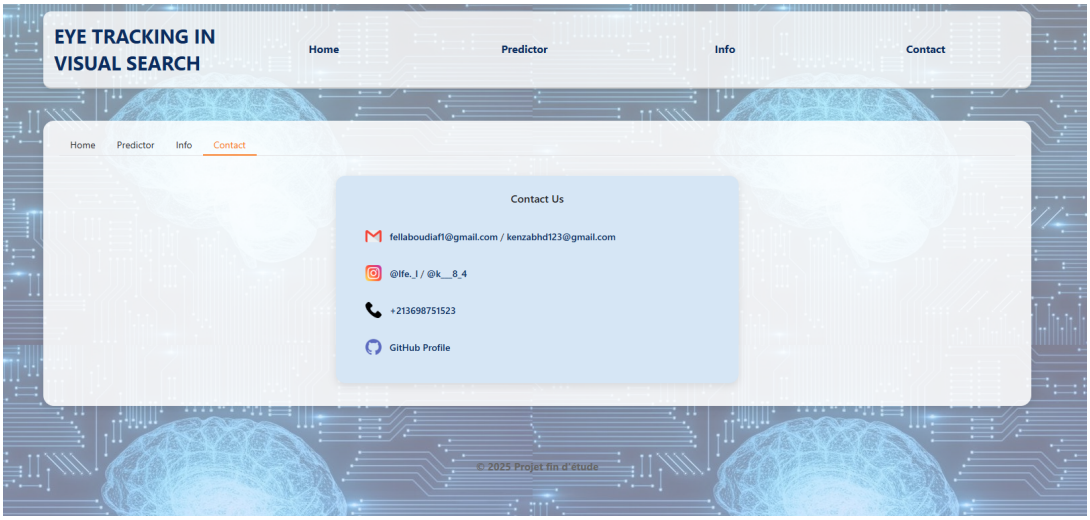


Figure 5.7: Contact Tab

Together, this structured interface streamlines the user journey from uploading an image to interpreting results making advanced visual attention prediction accessible to researchers, developers, and practitioners alike.

5.4 Results

To assess the effectiveness of the proposed eye-tracking system, the trained U-Net model was evaluated using a designated subset of the UEyes dataset. Its predicted saliency maps were compared against the ground truth using a combination of traditional error metrics and specialized saliency evaluation measures.

The performance of the model was quantified using the previous mentionned metrics:

MAE (Mean Absolute Error),MSE (Mean Squared Error),KLD (Kullback-Leibler Divergence),CC (Pearson's Correlation Coefficient),SIM (Similarity) and Accuracy

The table below summarizes the average performance of the model on five randomly selected test images:

Table 5.2: Statistical Summary of Model Performance

Metric	**Mean**	**Standard Deviation**
MAE	0.0469	$\pm$ 0.0081
MSE	0.0092	$\pm$ 0.0030
KLD	0.0321	$\pm$ 0.1793
CC	0.6510	$\pm$ 0.1726
SIM	0.5585	$\pm$ 0.0953
Accuracy	0.9762	$\pm$ 0.0167

In addition to the quantitative results, we presents a visual comparison of the original test images, their corresponding ground truth saliency maps, and the predicted saliency maps produced by the model. These qualitative visualizations demonstrate the model's ability to accurately capture prominent attention regions with reasonable spatial fidelity.

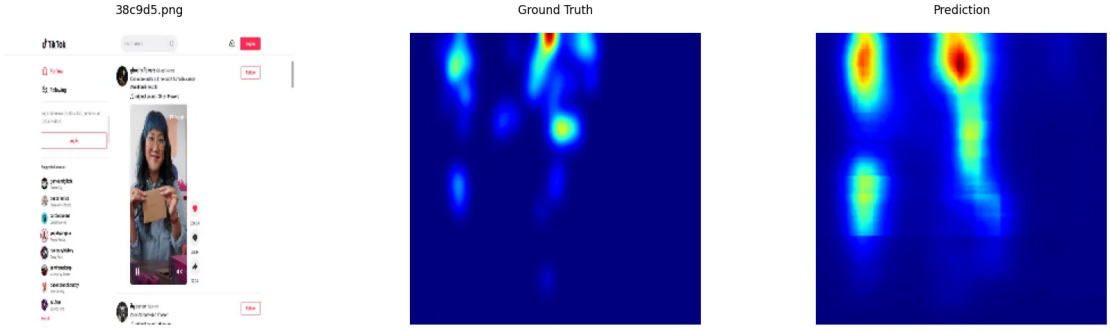
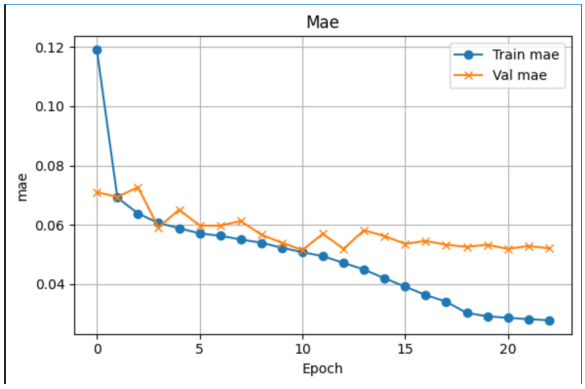
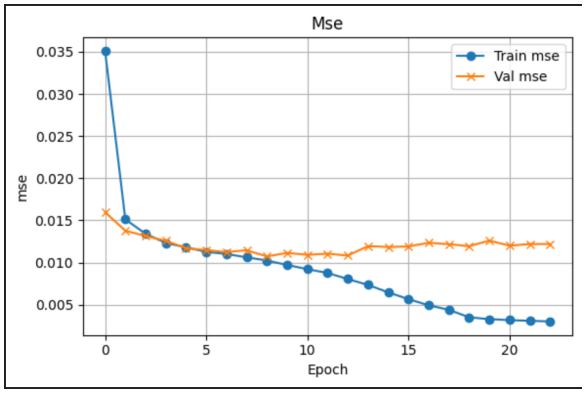


Figure 5.8: Visual comparison

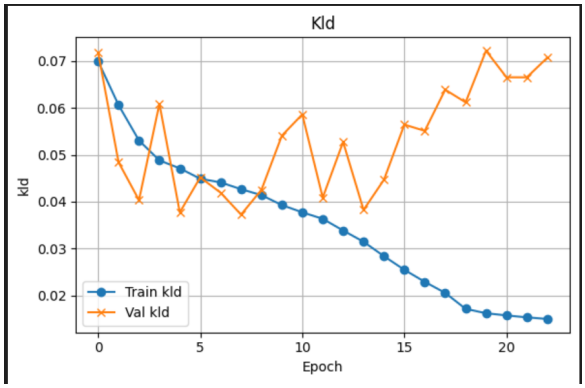
Moreover, to assess the the model's performance evolves and stabilizes during training, we plotted the training and validation graphs for each evaluation metric over epochs :



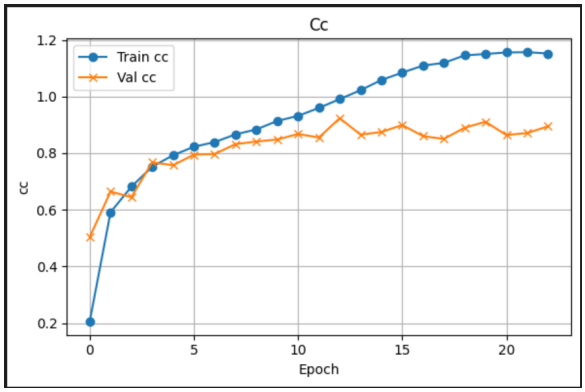
(a) MAE (Mean Absolute Error) graph



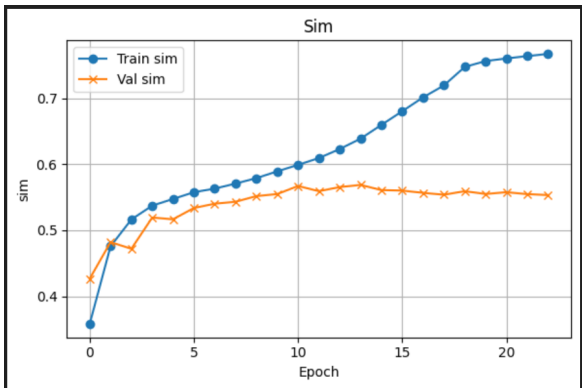
(b) MSE (Mean Squared Error) graph



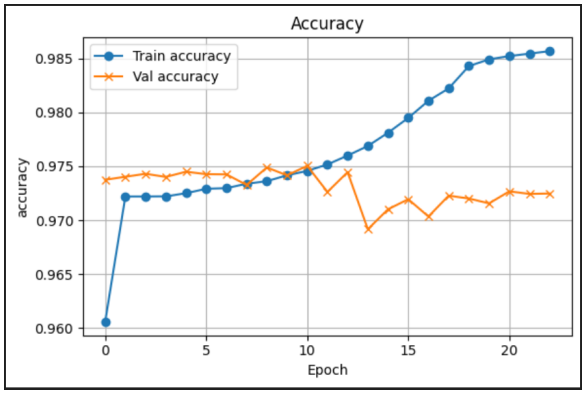
(c) KLD (Kullback-Leibler Divergence) graph



(d) CC (Pearson's Correlation Coefficient) graph



(e) SIM (Similarity) graph



(f) Accuracy graph

Figure 5.9: Training Vs Validation graphs of evaluation metrics over epochs

5.5 Discussion

The performance evaluation of the model provides a comprehensive view of its learning dynamics and generalization ability across training and validation phases. The training data, used to optimize the model's parameters, shows a steady improvement in performance . . As

seen in the Mean Absolute Error (MAE) curve: the training MAE decreases significantly from around 0.12 to below 0.03 over 22 epochs. The validation data, which the model does not see during training, exhibits a relatively stable MAE curve that hovers around 0.05. This close alignment between training and validation MAE implies the model is learning effectively with minimal overfitting, as there is no significant divergence between the two curves.

Complementing this, the Mean Squared Error (MSE) on the test set is 0.0092 ± 0.0030, further affirming the model's accuracy. Given that MSE penalizes larger errors more heavily than MAE, the low value here suggests rare and minimal large prediction errors. The combination of low MAE and MSE demonstrates that the model's predictions are both precise and stable.

The Kullback-Leibler Divergence (KLD) curve offers insight into how well the model's predicted saliency distributions match the ground truth. While the training KLD steadily decreases (indicating increasing alignment)the validation KLD shows more fluctuation and begins trending upward after epoch 10, suggesting a degree of overfitting as the model becomes more tailored to the training data. However, the overall test KLD remains reasonably low at 0.0321 ± 0.1793, indicating that the model still generalizes well across most test samples.

Turning to the Similarity (SIM) score, which assesses how structurally similar the predicted and actual saliency maps are, the model achieves a mean score of 0.5585 ± 0.0953. Since SIM values range from 0 (no overlap) to 1 (perfect match), a score above 0.55 reflects strong resemblance to human gaze patterns and highlights the model's capability to capture global visual attention distributions. The relatively low standard deviation further suggests stable and consistent predictions across different inputs.

The accuracy curve, measuring the binary correspondence between predicted and ground truth saliency masks, adds to the evidence of robust performance. Training accuracy climbs to around 98.5%, while validation accuracy remains steady at approximately 97.5%, showing minimal discrepancy between the two and indicating that the model generalizes well to unseen data without significant overfitting.

In conclusion, the convergence of training and validation metrics (MAE, Accuracy), the low error rates (MSE, KLD), and the meaningful SIM score collectively demonstrate that the U-Net model is highly effective and reliable for saliency prediction in eye-tracking tasks. The slight rise in validation KLD warrants consideration for regularization techniques like early stopping,

but overall, the model displays strong generalization, stability, and relevance to real-world eye movement analysis.

5.6 Comparative analysis with related work

To evaluate the performance of the proposed deep learning based eye-tracking system for simple visual search tasks, we compare our results against several state of the art models. The evaluation is based on standard saliency prediction metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Kullback–Leibler Divergence (KLD), Correlation Coefficient (CC), Similarity (SIM), and Accuracy.

Table 5.3: Comparison of performance measures for our proposed Eye-Tracking System with related systems

Study	**Method**	**Dataset**	**MAE ↓**	**MSE ↓**	**KLD ↓**	**CC ↑**	**SIM ↑**	**Accuracy ↑**
Kümmerer et al. (2015) [51]	DeepGaze I	MIT	0.078	0.016	1.12	0.70	0.56	-
Borji & Itti (2015) [52]	Classic Saliency Fusion	FIFA	0.082	0.021	0.87	0.60	0.51	-
Tavakoli et al. (2017) [53]	Bayesian Surprise Model	OSIE	0.075	0.018	0.62	0.61	0.53	-
Chen et al. (2021) [54]	ResNet + GRU	COCO-Search18	0.069	0.015	0.48	0.72	0.59	0.932
Zhang et al. (2022)[55]	Transformer Attention Tracker	VSCS	0.064	0.013	0.44	0.68	0.55	0.948
Che et al. (2023) [56]	SAM++, UMSI++ (ConvL-STM)	UEyes	0.052	0.011	0.045	0.640	0.543	0.961
Proposed	U-Net Saliency Model	**UEyes**	**0.0469**	**0.0092**	**0.0321**	**0.6510**	**0.5585**	**0.9762**

The comparative results presented for previous systems were taken directly from their published papers. We did not re-implement or rerun these models ourselves.

As shown in Table 5.3, our proposed U-Net-based model achieves superior performance across nearly all evaluation metrics, especially in MAE (0.0469), MSE (0.0092), KLD (0.0321), and Accuracy (0.9762). These improvements demonstrate the effectiveness of our architecture

in modeling goal-directed attention within user interfaces.

Compared to earlier models:

Kümmerer et al. (2015), using DeepGaze I on the MIT dataset, report a high KLD (1.12), likely due to the diversity and unconstrained nature of free-viewing tasks in their dataset.

Borji Itti (2015) and Tavakoli et al. (2017), evaluated on the FIFA and OSIE datasets respectively, show higher MAE and lower CC/SIM scores, highlighting limitations when dealing with semantically rich or poorly controlled environments.

Chen et al. (2021), using a ResNet+GRU on COCO-Search18, achieve a strong CC (0.72), but their higher MAE (0.069) and lower Accuracy (0.932) indicate challenges in modeling naturalistic search scenes.

Zhang et al. (2022), with a Transformer-based model on VSCS, show competitive CC and SIM values but still lag behind in MAE compared to our system.

Importantly, our model outperforms Che et al. (2023)—the closest prior work using the same UEyes dataset—across MAE, MSE, and Accuracy, confirming the strength of our U-Net architecture in UI-based saliency prediction tasks.

5.7 Conclusion

This chapter presented the U-Net-based eye-tracking system, highlighting its implementation, evaluation on the UEyes dataset, and interface via Gradio. The model achieved high accuracy and low error rates across multiple metrics, showed stable training behavior, and outperformed related methods, confirming its effectiveness for saliency prediction in UI-based visual search tasks.

Chapter 6

General Conclusion

6.1 Contributions

This project successfully designed and implemented a deep learning-based eye-tracking system for visual saliency prediction, utilizing the U-Net architecture and the uEyes dataset. A key contribution lies in the development of a U-Net model trained to produce saliency maps that effectively replicate human visual attention patterns. The model's performance was thoroughly evaluated using a suite of metrics demonstrating robust generalization across diverse inputs and a user-friendly Gradio interface was also developed, enabling real-time interaction and visualization of predicted saliency maps. Beyond the technical achievements, the project offers significant educational value by providing an in-depth exploration of eye-tracking principles, saliency modeling, and deep learning techniques. Overall, this work presents a comprehensive and adaptable framework for saliency prediction with potential applications

6.2 Limitations

Despite encouraging results, this work has several notable limitations. The evaluation is confined to static images, overlooking the temporal dynamics inherent in video sequences or dynamic scenes where gaze behavior changes continuously and requires models that can capture these temporal patterns.

Furthermore, the study focuses exclusively on traditional saliency maps, without investi-

gating alternative types of maps or methods that could provide richer or more comprehensive representations of visual attention. Examples include motion-based saliency maps or multi-channel attention maps that integrate both spatial and temporal information.

Additionally, the reliance solely on the U-Net architecture, while effective, may limit the model's potential, as no advanced architectural optimizations or combinations with other methods were explored to enhance prediction accuracy and robustness.

These limitations collectively affect the model's capability to generalize across more complex and diverse eye-tracking scenarios. Ethical Conditions This study relies on eye-tracking data derived from human participants, which inherently requires careful ethical consideration. Although the UEyes dataset used in this work was collected and publicly released by the original authors, it is important to acknowledge the ethical protocols that should be in place when conducting similar research involving human subjects.

6.3 Ethical Conditions

This study relies on eye-tracking data derived from human participants, which inherently requires careful ethical consideration. Although the UEyes dataset used in this work was collected and publicly released by the original authors, it is important to acknowledge the ethical protocols that should be in place when conducting similar research involving human subjects.The key ethical conditions include:

Informed Consent: All participants involved in the original data collection must have been clearly informed about the nature, purpose, and use of the eye-tracking study, and must have voluntarily agreed to participate.

Privacy and Anonymity: The dataset must not include any personal or identifiable information. Participant data should be anonymized to ensure individual privacy is protected.

Right to Withdraw: Participants must have had the right to withdraw from the study at any point without consequences.

Minimization of Harm: The eye-tracking sessions must be conducted in a safe, non-invasive manner that avoids causing physical or psychological discomfort.

Data Handling: The collected data must be stored securely and used only for research purposes. Ethical handling also includes avoiding any misuse or misinterpretation of gaze data, which could otherwise lead to incorrect profiling or conclusions.

Institutional Approval: Any research involving human participants typically requires approval from an ethics committee or institutional review board (IRB). While this project reused a publicly available dataset (UEyes), any future data collection should be preceded by appropriate ethical clearance.

These ethical conditions are critical to ensure the dignity, rights, and well-being of participants are protected, and that the research aligns with international ethical standards such as the Declaration of Helsinki or institutional ethical guidelines.

Since this project used an existing and publicly available dataset (UEyes), the responsibility for ethical compliance during data collection lies with the original researchers.

6.4 Future work and perspectives

Future work and several directions can be pursued to enhance and expand the eye-tracking system. A primary avenue involves extending the model to handle temporal dynamics by incorporating architectures suited for sequential data, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), temporal convolutional networks (TCNs), or transformer-based models. This would enable accurate saliency prediction in video sequences and dynamic environments, reflecting real-world gaze behavior more effectively.

Another important aspect is exploring diverse types of saliency and attention maps beyond traditional static saliency, including motion-based saliency maps and multi-channel maps that integrate spatial, temporal, and contextual information. This could provide richer representations and improve prediction robustness.

From a modeling perspective, applying advanced optimization strategies such as neural architecture search (NAS), Bayesian hyperparameter tuning, and ensemble learning could uncover more effective model architectures and training regimes. Integrating multimodal data—such as depth information, scene semantics, or user context—could also enhance saliency prediction accuracy.

Moreover, expanding the dataset to include a wider variety of participants, stimuli, and viewing conditions would improve model generalizability. Developing more sophisticated evaluation protocols that measure temporal consistency and task-specific relevance would provide deeper insights into model performance.

Finally, enhancing the usability and functionality of the system through more interactive interfaces and real-time feedback could facilitate broader adoption in applications ranging from human-computer interaction and advertising to medical diagnostics and cognitive research.

Bibliography

[1] A. T. Duchowski, *EyeTrackingMethodology*, 3rd ed. Andrew T. Duchowski School of Computing Clemson University Clemson, SC USA, 2017.

[2] A. Hollingworth and B. Bahle, "Eye tracking in visual search experiments," 2019. [Online]. Available: https://www2.psychology.uiowa.edu/faculty/hollingworth/documents/Holl_Bahle_Neuromethods.pdf

[3] Hessels et al., "The fundamentals of eye tracking part 1: The link between theory and research question," *Behavior Research Methods*, 2024. [Online]. Available: https://doi.org/10.3758/s13428-024-02544-8

[4] T. van Gog and K. Scheiter, "Eye tracking as a tool to study and enhance multimedia learning," *Learning and Instruction*, vol. 20, no. 2, pp. 95–99, 2010.

[5] V. Cantoni, C. Galdi, M. Nappi, M. Porta, and D. Riccio, "Eye tracking for the evaluation of human behavior in forensic and security applications," *Pattern Recognition Letters*, vol. 82, pp. 192–199, 2016.

[6] Y. Chen, Z. Yang, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky, "Coco-search18 fixation dataset for predicting goal-directed attention control," 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-87715-9

[7] K. Boyd and D. Turbert. (2023, Apr.) Eye anatomy: Parts of the eye outside the eyeball. American Academy of Ophthalmology. Reviewed by Ninel Z. Gregori, MD. [Online]. Available: https://www.aao.org/eye-health/anatomy/parts-of-eye

[8] E. R. Group, "human-eye-anatomy," https://www.elmanretina.com/services/human-eye-anatomy/, accessed: 2025-01-30.

[9] "Haptic-feedback-to-gaze-events scientific figure on researchgate," https://www.researchgate.net/figure/Example-of-Electro-OculoGraphy-EOG-Eye-movement-measurement-Picture_fig5_334988473, accessed: 30 jan 2025.

[10] A. Khaldi, E. Daniel, L. Massin, C. Kärnfelt, F. Ferranti, C. Lahuec, F. Seguin, V. Nourrit, and J.-L. de Bougrenet de la Tocnaye, "A laser emitting contact lens for eye tracking," *Scientific Reports*, vol. 10, 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-71233-1

[11] uav1, "infrared-(ir)-illuminator," https://www.uav1.com/wp-content/uploads/2023/06/230606-apple-vision-pro-wwdc23-12.jpg, accessed: 2025-01-30.

[12] Brigham Young University, https://brightspotcdn.byu.edu/dims4/default/d77fe9c/2147483647/strip/true/crop/386x350+0+0/resize/386x350!/quality/90/?url=https%3A%2F%2Fbrigham-young-brightspot-us-east-2.s3.us-east-2.amazonaws.com%2F41%2Ff7%2F56f78cb64f658ffcac5e8e894c73%2Feye-small.png, accessed: 2025-01-30.

[13] MDPI Vision Journal, https://www.mdpi.com/vision/vision-04-00025/article_deploy/html/images/vision-04-00025-g001.png, accessed: 2025-01-30.

[14] Western Sydney University, https://www.westernsydney.edu.au/__data/assets/image/0011/1930583/Eye_Tracker_Lab_-_LB_IMG_3997_-_reduced_size.JPG, accessed: 2025-01-30.

[15] Create CDN, https://sites.create-cdn.net/siteimages/74/1/0/741020/21/3/5/21357550/2000x1384.png?1738574647, accessed: 2025-04-30.

[16] Medium, https://miro.medium.com/v2/resize:fit:1100/format:webp/0*ibuldEl30zt-zDJN.png, accessed: 2025-05-30.

[17] kexxu, https://kexxu.com/?v=0de7b6a61a70, note = Accessed: 2025-04-30.

[18] R.-K. and colleagues, "Experimental design with three-camera-based-tracking-system," https://www.researchgate.net/publication/370127473/figure/fig1/AS:11431281151214464@1681959817632/a-Experimental-design-consists-of-three-camera-based-tracking-system-which-record-the.png, 2023, accessed: 2025-04-30.

[19] iMotions, "10 terms and metrics you should know for eye tracking research," https://imotions.com/blog/learning/10-terms-metrics-eye-tracking/, 2021, accessed: 2025-04-30.

[20] E. B. Huey, *The Psychology and Pedagogy of Reading*. Macmillan, 1908.

[21] R. Dodge and T. S. Cline, "The angle velocity of eye movements," *Psychological Review*, vol. 8, no. 2, pp. 145–157, 1901.

[22] G. T. Buswell, *How People Look at Pictures: A Study of the Psychology and Perception in Art*. University of Chicago Press, 1935.

[23] L. R. Young and D. Sheena, *Survey of eye movement recording methods*, 1975, vol. 7, no. 5.

[24] A. L. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.

[25] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension," *Psychological Review*, vol. 87, no. 4, pp. 329–354, 1980.

[26] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998.

[27] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011.

[28] H. Bogunović, F. G. Venhuizen, S. Klimscha *et al.*, "Prediction of anti-vegf treatment requirements in neovascular amd using a machine learning approach," *Scientific Reports*, vol. 7, p. 40564, 2017.

[29] M. Christopher, A. Belghith, C. Bowd *et al.*, "Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs," *Scientific Reports*, vol. 8, p. 16685, 2018.

[30] Z. Li, Y. He, S. Keel *et al.*, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 126, no. 8, pp. 1195–1204, 2019.

[31] S. Ghosh, S. Chanda, S. Chaudhary, and A. S. Jalal, "Gaze360: 360-degree gaze estimation via relational reasoning," https://arxiv.org/pdf/2309.06129, 2023, accessed: 2025-04-30.

[32] W. Fuhl, M. Tonsen, A. Bulling, and E. Kasneci, "Gazebase: A large-scale, multi-condition dataset for eye movement research," https://www.utdallas.edu/~nspain/GazeBase.html, 2017, university of Tübingen.

[33] Tobii Pro, tobii Pro Eye Tracking Data Sets. Tobii Technology. https://www.tobiipro.com/learn-and-support/learn/eye-tracking-datasets/.

[34] 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 (CVPR)*, 2015, https://vcai.mpi-inf.mpg.de/projects/mpiigaze/.

[35] P. Baheti, Y. Lan, J. Gazeau, J. M. Odobez, and D. Gatica-Perez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and depth data," https://www.idiap.ch/en/dataset/eyediap, 2014, idiap Research Institute.

[36] J. Kim and S. Lee, "Opengaze: A toolkit and dataset for gaze-based interaction," https://github.com/opengaze/OpenGaze, 2020, gitHub Repository.

[37] J. Chen, G. J. Zelinsky, and H. Adeli, "Coco-search18: A dataset for predicting goal-directed attention control," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, https://www.researchgate.net/publication/351064716_COCO-Search18_fixation_dataset_for_predicting_goal-directed_attention_control.

[38] Y. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, vol. 14, no. 1, pp. 28–28, 2014, oSIE Datasethttps://www-users.cse.umn.edu/~qzhao/predicting.html,.

[39] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, mIT Saliency Benchmarkhttp://saliency.mit.edu/.

[40] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear

manner," *Journal of Vision*, vol. 11, no. 3, p. 9, 2011.

[41] J. M. Wolfe and M. S. Cain, "Visual search in cluttered scenes: A dataset for testing human and model performance," *Journal of Vision*, vol. 18, no. 10, p. 9, 2018.

[42] Y. Jiang, L. A. Leiva, H. R. Tavakoli, P. R. B. Houssel, J. Kylmälä, and A. Oulasvirta, "UEyes: Understanding visual saliency across user interface types," New York, NY, USA, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3544548.3581096

[43] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[45] B. Dickson, "Demystifying deep learning," https://bdtechtalks.com/2021/01/28/deep-learning-explained/, 2021, accessed: 2025-05-07.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015, accessed: 2025-04-30. [Online]. Available: https://arxiv.org/abs/1505.04597

[47] V. Joshi, "U-net architecture — explained," https://www.geeksforgeeks.org/u-net-architecture-explained/, 2023, accessed: 2025-04-30.

[48] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[49] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[50] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[51] M. Kümmerer, L. Theis, and M. Bethge, "Deepgaze i: Boosting saliency prediction with

feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2015.

[52] A. Borji and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," *IEEE International Conference on Computer Vision (ICCV)*, pp. 921–928, 2015.

[53] H. Tavakoli, A. Borji, and J. Laaksonen, "Saliency revisited: Analysis of mouse movements versus fixations," *Proceedings of the IEEE CVPR*, pp. 1571–1579, 2017.

[54] T. Chen, G. J. Zelinsky, and E. Adeli, "Coco-search18 fixation dataset for goal-directed attention prediction," *Scientific Reports*, vol. 11, no. 1, p. 20613, 2021.

[55] H. Zhang, W. Lin, Q. Zhao, and Z.-H. Zhou, "Target-guided transformer for visual search in cluttered scenes," *Neural Networks*, vol. 153, pp. 389–400, 2022.

[56] W. Che, Y. Huang, W. Xu, X. Fu, and X. Lian, "Ueyes: A benchmark dataset for eye tracking in user interfaces," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023.