ISTANBUL TECHINICAL UNIVERSITY ★ GRADUATE SCHOOL

DEVELOPING A NEW SYSTEM FOR ADVERTISEMENT ANALYSIS USING GAZE AND DEPTH ANALYSIS METHODS

M.Sc. THESIS

Fatih BADAY

Department of Big Data and Business Analytics

Big Data and Business Analytics Programme

AUGUST 2024

ISTANBUL TECHINICAL UNIVERSITY ★ GRADUATE SCHOOL

DEVELOPING A NEW SYSTEM FOR ADVERTISEMENT ANALYSIS USING GAZE AND DEPTH ANALYSIS METHODS

M.Sc. THESIS

Fatih BADAY
(528211078)

Department of Big Data and Business Analytics

Big Data and Business Analytics Programme

Thesis Advisor: Assoc. Prof. Dr. Ömer Faruk BEYCA

AUGUST 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ ENSTİTÜSÜ

BAKIŞ VE DERİNLİK ANALİZİ YÖNTEMLERİ KULLANILARAK REKLAM ANALİZİ İÇİN YENİ BİR SİSTEM GELİŞTİRİLMESİ

YÜKSEK LİSANS TEZİ

Fatih BADAY
(528211078)

Büyük Veri ve İş Analitiği Anabilim Dalı

Büyük Veri ve İş Analitiği Programı

Tez Danışmanı: Doç. Dr. Ömer Faruk BEYCA

AĞUSTOS 2024

Fatih BADAY, a M.Sc. student of ITU Graduate School student ID 528211078, successfully defended the thesis entitled “DEVELOPING A NEW SYSTEM FOR ADVERTISEMENT ANALYSIS USING GAZE AND DEPTH ANALYSIS METHODS”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assoc. Prof. Dr. Ömer Faruk BEYCA** Istanbul Technical University	....
Jury Members :	**Prof. Dr. Selim ZAİM** IBN Haldun University	....
	Assoc. Prof Dr. Fuat KOSANOĞLU Istanbul Technical University	....

Date of Submission : 24 May 2024
Date of Defense : 15 August 2024

FOREWORD

This thesis represents the culmination of a significant journey of learning, discovery, and growth. I am deeply grateful to those who have supported and guided me throughout this process.

Firstly, I would like to express my sincere gratitude to my advisor, Doç. Dr. Ömer Faruk BEYCA, whose expertise, patience, and unwavering support have been instrumental in the completion of this thesis. Their insightful feedback and encouragement have been invaluable.

I extend my heartfelt thanks to the faculty and staff of the Big Data and Business Analytics Department at Istanbul Technical University. Their dedication to excellence in education and research has provided me with the knowledge and skills necessary to undertake this project.

I hope the findings and insights presented in this work will contribute to further research and practical applications in the field.

Thank you all for your support and encouragement.

August 2024

Fatih BADAY
(Data Scientist)

TABLE OF CONTENTS

Page

ABBREVIATIONS

CNN	**:** A convolutional neural network used in image processing.
LSTM	**:** A type of neural network used for processing sequential data.
ResNet	**:** A deep learning model that reduces information loss between layers.
AR	**:** A technology that overlays digital information onto the real world.
HCI	**:** A field of study focused on the interaction between humans and computers.
RPN	**:** A neural network used to suggest regions of interest for object detection in images.
RoIs	**:** Specific areas of interest within an image.
RGB	**:** A color model used in digital imaging.
UV	**:** A coordinate system used to map a 2D surface to a 3D model.
ReLU	**:** A mathematical function used as an activation function in neural networks.
FCN	**:** A type of neural network used for tasks such as image segmentation.

LIST OF FIGURES

DEVELOPING A NEW SYSTEM FOR ADVERTISEMENT ANALYSIS USING GAZE AND DEPTH ANALYSIS METHODS

SUMMARY

This thesis proposes an innovative methodology based on observers' perspectives to deeply understand the interaction and persuasive capacity of advertisements on individuals. The core aim of this study is to develop a system capable of assessing the interest in advertising content based on the directions of observers' gazes. In this thesis, four fundamental steps have been followed to thoroughly examine the interest towards advertising content. Each step involves the technological methods and algorithms of a system that contributes to understanding the interactive power of advertisements, starting from the observers' viewpoints. The research begins with the identification of human figures and their eye positions in video frames using the DensePose algorithm. Following this step, the Gaze360 model, which utilizes deep learning techniques to predict individuals' viewing angles, is introduced. Developed by a group of researchers at MIT, this model advances on the foundation of ResNet architectures to determine where individuals are looking. In the third stage, depth information obtained through the ManyDepth algorithm is used to ascertain the spatial positions of objects in images, and this information is employed to create a depth map. The integrated data are utilized to determine which advertising contents are more engaging to viewers. The study emphasizes the process of carefully adjusting and testing the integration of the technological systems used, with the goal of providing a detailed examination of interaction dynamics within the advertising sector. As a result, it develops new methods that allow for a more accurate measurement of the impact of advertisement content on viewers. Instead of focusing on specific field tests, this thesis explores the theoretical foundations of the proposed methodology and offers a comprehensive framework on how user interaction in the advertising sector can be more effectively optimized. This approach aims to lay a foundation for future research and to contribute methodological and theoretical insights towards enhancing the effectiveness of advertising strategies. Therefore, it seeks to aid in the development of more effective advertising techniques and to maximize the persuasive power of advertisements.

BAKIŞ VE DERİNLİK ANALİZİ YÖNTEMLERİ KULLANILARAK REKLAM ANALİZİ İÇİN YENİ BİR SİSTEM GELİŞTİRİLMESİ

ÖZET

Bu tez çalışması, reklamların bireyler üzerindeki etkileşim gücünü ve etkisini anlamaya yönelik yenilikçi bir metodoloji geliştirmeyi hedeflemektedir. Özellikle dijital reklamcılık dünyasında, reklamların izleyiciler üzerindeki dikkat çekme potansiyelinin doğru bir şekilde analiz edilmesi, reklam stratejilerinin etkinliğini artırmada kritik öneme sahiptir. Bu bağlamda tezde, göz izleme (gaze tracking) ve derinlik analizi (depth analysis) yöntemleri bir araya getirilerek, izleyicilerin bakış yönlerini ve mekânsal konumlarını kullanarak reklam içeriklerine olan ilginin değerlendirilmesi hedeflenmiştir. Günümüzde reklamcılık sektörü, dijitalleşme ile birlikte büyük bir hızla evrilmekte ve yeni yöntemlerle izleyicilerin dikkatini çekmeye çalışmaktadır. 2023 yılında küresel reklam harcamalarının yaklaşık 733 milyar dolara ulaştığı tahmin edilmekte olup, 2026 yılına kadar bu rakamın 800 milyar doları aşması beklenmektedir. Bu rakamlar, reklamcılığın ticari başarıda oynadığı kritik rolü açıkça ortaya koymaktadır. Ancak, bu büyük yatırımın karşılığını alabilmek için reklamların gerçekten izleyici üzerinde nasıl bir etki bıraktığını anlamak gerekmektedir. Geleneksel metrikler, izlenim sayısı (impressions), tıklama oranları (click-through rates) ve dönüşüm oranları gibi yüzeysel veriler sunmakta; ancak izleyicilerin reklamlarla olan gerçek etkileşim düzeyini tam olarak yansıtmamaktadır. Bu tez çalışması, izleyicilerin reklamlarla nasıl etkileşime geçtiğini daha derinlemesine analiz etmek için göz izleme ve derinlik analizi teknolojilerini bir araya getirerek, daha doğru ve kapsamlı sonuçlar elde etmeyi amaçlamaktadır. Bu tezin temel amacı, izleyicilerin reklamlara olan ilgisini daha iyi anlamak için bir sistem geliştirmektir. Bu amaç doğrultusunda dört temel adım izlenmiştir. İlk adım, video karelerinde izleyicilerin insan figürlerinin ve göz pozisyonlarının tespit edilmesidir. Bu işlem için kullanılan DensePose algoritması, video karelerindeki insan figürlerini 3D modeller üzerine haritalayarak, izleyicilerin bakış pozisyonlarını doğru bir şekilde tespit edebilmiştir. DensePose, insan vücudundaki her bir pikseli bir 3D insan vücudu modeli üzerinde haritalandırarak, izleyicilerin beden ve göz konumlarını yüksek hassasiyetle

belirlemektedir. Bu aşama, izleyicilerin reklamlarla nasıl etkileşimde bulunduğunu anlamak için kritik bir öneme sahiptir. İzleyicilerin reklamları izlerken hangi yönlere baktıkları ve beden duruşları, reklamların çekiciliği ve etkisi hakkında önemli ipuçları vermektedir. İkinci adımda, izleyicilerin bakış yönleri Gaze360 modeli kullanılarak tahmin edilmiştir. Gaze360, derin öğrenme teknikleri kullanarak bireylerin nereye baktığını tahmin edebilen gelişmiş bir modeldir. Gaze360, MIT'deki araştırmacılar tarafından geliştirilmiş olup, özellikle ResNet mimarisi üzerine inşa edilmiştir. Bu model, bireylerin göz hareketlerini doğru bir şekilde izleyerek, bakış yönlerini tahmin edebilmekte ve izleyicilerin nereye baktığını yüksek doğrulukla belirlemektedir. Geleneksel göz izleme yöntemlerinin aksine, Gaze360, kontrol edilmemiş ve doğal koşullarda da doğru sonuçlar verebilmektedir. Bu özellik, özellikle dış mekân reklamcılığı gibi alanlarda önemli avantajlar sunmaktadır. İzleyicilerin doğal bir şekilde reklamları izlerken bakışlarının izlenmesi, reklamın izleyici üzerinde ne kadar dikkat çekici olduğunu belirlemek açısından büyük önem taşımaktadır. Üçüncü adımda, izleyicilerin bulundukları ortamın mekânsal derinlik bilgileri ManyDepth algoritması kullanılarak elde edilmiştir. ManyDepth, ardışık iki görüntüden mekânsal derinlik bilgisi çıkarabilen, gelişmiş bir derinlik tahmin algoritmasıdır. Bu algoritma, görüntülerdeki nesnelerin mekânsal konumlarını belirlemek için kullanılmakta olup, derinlik haritaları oluşturarak izleyicilerin bulundukları ortamın üç boyutlu yapısını anlamaya yardımcı olmaktadır. Özellikle dış mekân reklamları söz konusu olduğunda, bir reklamın izleyiciye olan uzaklığı ya da yakınlığı, reklamın ne kadar dikkat çekici olabileceği ile doğrudan ilişkilidir. ManyDepth algoritması sayesinde izleyicilerin reklamlarla olan mekânsal ilişkisi doğru bir şekilde hesaplanabilmiştir. Bu sayede reklamların izleyici üzerindeki mekânsal etkisi analiz edilebilmiştir. Son aşamada, göz ve derinlik verileri birleştirilerek, izleyicilerin hangi reklam içeriklerine daha fazla ilgi gösterdiği belirlenmiştir. Reklam analizinde izleyicilerin bakış yönleri ve mekânsal konumlarının entegre edilmesi, reklamların izleyici üzerindeki etkisini daha doğru bir şekilde ölçmeye olanak tanımaktadır. Göz izleme ve derinlik bilgilerini birleştiren bu sistem, izleyicinin reklamı gerçekten görüp görmediğini anlamak için oldukça etkili bir yöntem sunmaktadır. Reklamın izleyicinin dikkatini çekip çekmediğini anlamak, reklam stratejilerinin başarısını doğrudan etkileyen bir faktördür. İzleyicilerin hangi reklam içeriklerine ne kadar süreyle dikkat ettiklerinin anlaşılması, reklam stratejilerinin daha etkili bir şekilde optimize edilmesine yardımcı olabilecektir. Tezde geliştirilen sistemin performansı, farklı testler aracılığıyla değerlendirilmiş ve başarılı sonuçlar elde edilmiştir. Yapılan deneylerde, izleyicilerin bakış yönleri, mekânsal konumları ve göz hareketleri detaylı bir şekilde analiz edilmiştir. Bu analizler sonucunda, reklamların izleyiciler üzerindeki gerçek etkisinin daha net bir şekilde anlaşılabileceği görülmüştür. Testler sırasında, her bir video karesi üzerinde sistemin dört ana algoritması (DensePose, Gaze360, ManyDepth) çalıştırılmış ve elde edilen veriler entegre edilerek izleyicilerin hangi reklamlara baktığı belirlenmiştir. Bu analizler sonucunda, reklam içeriklerinin izleyiciler üzerindeki etkisinin ölçülmesinde kullanılan geleneksel yöntemlerden daha ileri düzeyde sonuçlar elde edilmiştir. Ayrıca, geliştirilen sistemin gerçek zamanlı veri işleme kapasitesi de test edilmiştir. Sistem, saniyede 5 kare işleyebilme kapasitesine sahip olup, bu hız, dinamik reklam senaryolarında gerçek zamanlı analizler yapılmasına olanak tanımaktadır. Bu özellik, özellikle dijital dış mekân reklamcılığı gibi anlık geri bildirimin önemli olduğu alanlarda büyük bir avantaj sağlamaktadır. Reklamların izleyici üzerindeki anlık etkisinin ölçülmesi, reklam stratejilerinin anlık olarak optimize edilmesine yardımcı olabilecektir. Sonuç olarak, bu tez çalışması, hem reklamcılık sektörüne hem de bilgisayarla görü ve derin öğrenme alanlarına önemli katkılar sunmaktadır. Göz izleme ve derinlik analizine dayalı olarak geliştirilen bu sistem, reklam içeriklerinin izleyiciler üzerindeki etkisini daha doğru bir şekilde ölçmeyi sağlamaktadır. İzleyicilerin reklamlara ne kadar dikkat ettiklerini ve hangi unsurların dikkat çekici olduğunu anlamak, reklam stratejilerinin daha etkili bir şekilde oluşturulmasına yardımcı olacaktır. Bu bağlamda, geliştirilen sistemin gelecekteki araştırmalara ve pratik uygulamalara katkı sağlayacağı öngörülmektedir. Bu tez çalışması, izleyicilerin reklamlara olan ilgisini daha doğru bir şekilde analiz edebilmek için ileri teknoloji yöntemlerini bir araya getiren yenilikçi bir yaklaşımla, reklam analizinde önemli bir boşluğu doldurmayı başarmıştır. Göz izleme ve derinlik analizinin bir arada kullanıldığı bu sistem, izleyici ilgisini daha iyi anlamak ve reklam stratejilerini optimize etmek için güçlü bir temel sunmaktadır. Gelecekte, bu sistemin daha geniş veri setleri ile test edilmesi ve farklı reklam formatlarına genelleştirilmesi planlanmaktadır.

1. INTRODUCTION

1.1 Context and Background

Advertising is a pivotal component of the global economy, influencing consumer behavior and driving market trends. According to recent reports, the global advertising market was valued at approximately $733 billion in 2023 and is projected to increase to over $800 billion by 2026, indicating the sector's significant economic impact [1]. This growth underscores the critical role advertising plays in commercial success across industries.

In the digital realm, advertising expenditures are not only robust but also growing significantly. For instance, worldwide digital advertising spending reached $549.51 billion in 2022. It is projected to increase further, with expectations to reach $870.85 billion by the year 2027 [2]. Despite substantial investments in marketing, the digitalization of consumer media consumption still faces significant challenges. Contrary to often cited figures, a recent analysis suggests that an adult in a metropolitan area sees between 50 to 400 ads per day across various platforms, including online, TV, and billboards. This range varies widely based on location, lifestyle, and consumption habits. Importantly, not all ads are consciously noticed by individuals [3]. This saturation complicates advertisers' efforts to capture attention and significantly diminishes the impact of individual advertisements.

Traditional metrics such as impressions, click-through rates, and conversion rates remain commonly used to gauge advertisement effectiveness. However, these metrics often fail to provide a complete picture of engagement. For example, a high number of impressions indicates extensive exposure but does not necessarily reflect active viewer engagement or attention. This gap highlights the limitations of conventional analytical tools and underscores the necessity for more advanced methodologies capable of offering deeper insights into consumer engagement.

Emerging technologies such as eye-tracking and depth analysis offer new opportunities to enhance advertisement analysis. Originally utilized in psychological studies, eye-tracking has evolved into a crucial tool for assessing where and how viewers direct their attention in real-time [4]. When combined with depth analysis, which evaluates interactions within three-dimensional spaces, these technologies can reveal complex patterns of viewer engagement that traditional metrics overlook. By leveraging these advanced technologies, advertisers can gain a more comprehensive understanding of how advertisements are perceived and interacted with, thus enabling more effective and engaging advertising strategies.

1.2 Theoretical Foundation

The theoretical underpinnings of this thesis are rooted in the interdisciplinary fields of cognitive psychology, marketing, and computer vision, each contributing essential perspectives to the analysis of advertisement engagement.

1.2.1 Cognitive attention in adversiting

The concept of the Attention Economy is fundamental in understanding how and where consumers allocate their attention in the modern information age. This theory emphasizes that human attention is especially valuable in environments rich in information and saturated with advertisements [5,6]. Research in cognitive psychology indicates that attention serves as the gateway to memory and decision-making processes, significantly influencing consumer behavior [7]. Eye-tracking technology provides a direct measure of visual attention by recording eye movements, which are closely linked to attentional focus [8].

1.2.2 Integration of gaze and depth analysis

The integration of gaze and depth analysis in advertising research is a relatively novel approach but is supported by advancements in computer vision and machine learning. These technologies provide a framework for understanding not just the focus of attention but also the context in which that attention occurs [9]. For instance, studies utilizing 3D environments and virtual reality have shown that depth information significantly influences user experience and engagement [10].

1.2.3 Empirical evidence in advertising research

Empirical studies have demonstrated the efficacy of using eye-tracking for advertisement analysis. Further, depth analysis has been increasingly applied in studies examining consumer behavior in virtual stores and digital interfaces, highlighting its potential to enhance traditional advertising metrics.

By grounding the study in these theoretical frameworks, this research not only taps into well-established scientific principles but also pioneers the application of these theories in a combined gaze and depth analytical approach. The intention is to provide a more robust and nuanced understanding of how advertisements engage viewers, extending beyond traditional metrics to include physiological and psychological dimensions of viewer interaction.

1.3 Problem Statement

Accurately measuring the true impact and effectiveness of outdoor advertising campaigns remains a significant challenge. Traditional metrics used in digital advertising—such as impressions, click-through rates, and conversion rates—provide valuable insights but fall short in the context of outdoor ads. These metrics do not offer a nuanced understanding of consumer engagement with physical advertisements.

Current methods for evaluating outdoor advertisement engagement, such as traffic counts and self-reported surveys, fail to precisely measure actual viewer engagement. These methods are often inadequate in dynamic environments where factors like pedestrian movement and varying weather conditions influence ad reception. Consequently, they do not accurately reflect how many people truly view and engage with the advertisements.

Traditional approaches also do not account for the complex and non-linear nature of human gaze behavior in naturalistic settings. While eye-tracking technology has been used to map where viewers look in controlled environments, integrating this data to understand how consumers perceive and interact with outdoor advertisements in real-world, three-dimensional spaces remains underexplored. This gap is significant because depth perception and the surrounding environmental context play crucial roles in how visual information is processed.

Therefore, there is a need for a new methodological approach that not only tracks where a viewer's gaze lands but also contextualizes this within the spatial dynamics of the outdoor advertising environment. This thesis proposes a novel approach that leverages a combination of deep learning algorithms to track and analyze where people are looking. By providing a more accurate count and understanding of engaged viewers, this method aims to offer a clearer and more comprehensive assessment of the true impact of outdoor advertisements.

Such an approach would significantly enhance our understanding of consumer behavior and improve the evaluation of outdoor advertising effectiveness, ultimately helping advertisers better understand and quantify the real engagement of their audience.

2. LITERATURE REVIEW

The effectiveness of advertisements has been extensively studied, integrating insights from various disciplines such as marketing, psychology, and communication studies. With technological advancements in computer vision and deep learning, new methodologies have emerged to measure and understand advertising effectiveness from the observer's perspective. This literature review delves into these methodologies, focusing on key technologies such as DensePose for human pose estimation, Gaze360 for gaze direction prediction, and ManyDepth for depth information extraction. These cutting-edge techniques provide a sophisticated approach to analyzing visual attention and engagement with advertisements, offering valuable insights into viewer interaction dynamics.

2.1 Advertising Effectiveness and Visual Attention

Role of Visual Attention in Advertising

Visual attention is a crucial determinant of advertising effectiveness. It involves the selective focus on specific elements within an advertisement, significantly influencing the viewer's ability to process and recall the advertisement's message. Research has consistently shown a strong correlation between visual attention and key advertising outcomes, such as brand recall, attitude towards the ad, and purchase intention [11]. Visual attention is often influenced by various factors including the layout, color, contrast, and overall design of the advertisement.

Eye-Tracking Technology in Advertising Research

Eye-tracking technology has been fundamental in measuring visual attention within advertising research. Eye-tracking devices record the viewer's gaze points and movements, providing detailed data on how viewers interact with different elements of an advertisement. This data helps researchers understand which parts of an ad capture the most attention and how viewers process visual information [12]. Eye-tracking studies have uncovered critical patterns such as the prominence of visual hierarchy and the impact of color and contrast in capturing attention [13]. Additionally, eye-tracking data can be used to identify fixation points, saccades, and gaze paths, offering a granular view of viewer engagement.

2.2 Human Pose Estimation and DensePose

Advancements in Human Pose Estimation

Human pose estimation is a foundational task in computer vision, essential for understanding human activities and interactions within visual data. DensePose, developed by Güler, Neverova, and Kokkinos, represents a significant advancement in this field [14]. DensePose maps all human pixels of an RGB image to the surface of a 3D human body model, providing precise and detailed information about body positions and orientations.

Methodology and Applications of DensePose

DensePose employs deep convolutional neural networks (CNNs) to achieve high accuracy in human pose estimation. It leverages extensive datasets with dense annotations, enabling it to map image pixels to 3D body models accurately. This capability is particularly beneficial for advertising research, as it allows for the precise detection of viewers' body and eye positions, which are crucial for accurate gaze estimation [14]. DensePose's methodology involves training on large-scale datasets with densely annotated human poses, ensuring high precision and robustness in real-world scenarios.

Impact on Advertising Research

The application of DensePose in advertising research offers new dimensions of understanding viewer engagement. By accurately estimating the pose and orientation of viewers, researchers can analyze how different advertisements capture attention based on body language and orientation. For instance, understanding body orientation and movement in response to advertisements can provide insights into viewer engagement and the effectiveness of ad placements in physical environments such as billboards or digital signage in public spaces.

2.3 Gaze Estimation with Gaze360

Importance of Gaze Estimation

Gaze estimation involves predicting the direction in which a person is looking, which is essential for understanding visual attention and engagement with advertising content. Accurate gaze estimation helps identify which elements of an advertisement draw the viewer's attention and how effectively the advertisement communicates its message.

Gaze360 Model and Its Applications

Gaze360 offers an innovative solution for predicting gaze direction from unconstrained images [15]. Gaze360 uses a deep learning approach based on spherical regression to estimate gaze vectors, allowing for accurate gaze estimation even in challenging real-world scenarios. The Gaze360 model is based on the ResNet architecture, a highly effective deep neural network known for its residual learning capabilities. This architecture enables the model to handle large variations in head pose and environmental conditions, making it suitable for applications in advertising research where viewers may be observed in diverse settings [15].

Gaze Estimation in Dynamic Environments

One of the significant advantages of Gaze360 is its ability to estimate gaze direction in dynamic environments. Unlike traditional eye-tracking methods that require controlled settings, Gaze360 can be applied in naturalistic environments, providing more ecologically valid data. This feature is particularly useful for studying outdoor advertising and digital billboards where viewer movement and varying environmental conditions are common.

Enhancing Advertising Research with Gaze360

The application of Gaze360 in advertising research allows for a more nuanced understanding of how viewers engage with advertisements. By accurately estimating gaze direction, researchers can determine which elements of an advertisement attract the most attention and how this attention varies across different contexts and environments. This information is crucial for designing advertisements that maximize viewer engagement and effectiveness.

2.4 Depth Estimation and ManyDepth Algorithm

Significance of Depth Estimation

Depth estimation is the process of determining the distance between objects in a scene and the camera. This information is vital for understanding the spatial relationships between objects and the viewer's gaze. Depth information helps in creating a three-dimensional understanding of the scene, which is crucial for analyzing viewer engagement with advertisements.

ManyDepth Algorithm and Its Impact

ManyDepth, an advanced algorithm for monocular depth estimation, uses self-supervised learning to achieve high accuracy without the need for ground-truth depth annotations [16]. ManyDepth predicts disparity maps from single images, which are then used to generate depth maps representing the scene's spatial structure. This depth information is invaluable in advertising research, as it allows for the analysis of how spatial arrangements of elements within an ad influence viewer engagement. Understanding the depth and spatial context of advertisements can help in designing more effective visual content [16].

Application in Advertising Layouts

The application of depth estimation in advertising research can significantly impact how advertisements are designed and placed. By understanding the spatial relationships within an advertisement, researchers can identify optimal placements for key visual elements to maximize viewer engagement. For example, elements that appear closer to the viewer might be more attention-grabbing, suggesting that depth cues can be strategically used to enhance the effectiveness of advertisements.

Enhancing Viewer Interaction with Depth Information

Integrating depth information into advertising design can also enhance viewer interaction by creating more immersive and engaging experiences. For instance, three-dimensional effects and spatial arrangements that mimic real-world depth can make advertisements more visually appealing and engaging, leading to higher levels of viewer retention and impact.

2.5 Integration of Visual Attention and Depth Information

Combining Gaze and Depth Data

The integration of visual attention data and depth information provides a comprehensive approach to analyzing advertising effectiveness. By combining gaze estimation with depth maps, researchers can gain insights into how viewers interact with three-dimensional advertising spaces. This integrated approach enables the identification of elements that capture attention in different spatial contexts, thereby enhancing the understanding of how ad placement and design affect viewer engagement.

Technological Integration in Real-Time Advertising

Real-time applications of this integrated approach can revolutionize digital advertising. For instance, interactive billboards that adjust their content based on real-time gaze and depth data can provide personalized advertising experiences. Such technologies can enhance engagement by dynamically responding to viewer behavior, creating more effective and immersive advertising environments.

Enhancing Advertising Strategies with Integrated Data

The integration of visual attention and depth data can significantly enhance advertising strategies by providing a more holistic understanding of viewer engagement. Advertisers can leverage this information to design more effective advertisements that capture and retain attention, ultimately leading to better advertising outcomes.

2.6 Technological Advances in Advertising Research

Advancements in Computer Vision

The use of advanced computer vision and deep learning techniques in advertising research represents a significant technological advancement. Traditional methods such as self-reports and focus groups often suffer from biases and limitations in capturing real-time attention and engagement. In contrast, the use of DensePose, Gaze360, and ManyDepth algorithms provides objective, precise, and data-driven insights into viewer behavior and engagement with advertisements.

Innovations in Deep Learning Models

Deep learning models, particularly those based on convolutional neural networks (CNNs) and residual networks (ResNets), have revolutionized the analysis of visual data. These models can handle complex patterns and large variations in data, making them ideal for applications in advertising research. Innovations such as transfer learning and self-supervised learning further enhance the capabilities of these models, allowing them to learn from limited labeled data and adapt to new scenarios [17].

3. METHODOLOGY

This chapter outlines the comprehensive methodologies adopted to analyze viewer interactions with outdoor advertisements, focusing on the integration of advanced gaze estimation, human pose estimation, and depth analysis technologies. The objective is to correlate these interactions with the effectiveness of advertising strategies, providing insights into consumer behavior and engagement.

3.1 Densepose

DensePose is a sophisticated computer vision algorithm designed to establish dense correspondences between 2D RGB images and a 3D surface model of the human body. This capability is essential for understanding human pose and body parts in a detailed manner, which is crucial for applications such as augmented reality (AR), human-computer interaction (HCI), and in our case, analyzing gaze direction to assess the effectiveness of advertisements.

Developed by Güler, Neverova, and Kokkinos, DensePose was introduced in their paper "DensePose: Dense Human Pose Estimation In The Wild" [14]. The primary objective of DensePose is to map every pixel of a human body in a 2D image to its corresponding point on a 3D human body model. This mapping facilitates precise localization and tracking of body parts, enabling a comprehensive analysis of human interactions within images.

DensePose employs a convolutional neural network (CNN) architecture, leveraging a ResNet-based backbone for feature extraction. ResNet, known for its ability to learn residual representations, is instrumental in mitigating the vanishing gradient problem often encountered in deep networks, thereby enhancing learning efficiency and performance.

The DensePose model architecture comprises several key components:

1. ResNet Backbone: Processes the input image to generate high-level feature maps. ResNet (Residual Network) is chosen for its robust performance in various image recognition tasks.
2. Region Proposal Network (RPN): Identifies regions of interest (RoIs) within the feature maps that likely contain human figures, focusing subsequent processing on specific areas of the image.
3. RoIAlign Layer: Extracts fixed-size feature maps from the identified RoIs, ensuring spatial alignment preservation, which is crucial for accurate mapping and pose estimation.
4. DensePose Head: Consists of fully convolutional layers that predict dense correspondences between the RoI feature maps and the UV coordinates on the 3D human body model.

The DensePose pipeline involves several steps to achieve dense human pose estimation:

1. Input Image Processing: An RGB image containing human figures is input into the system.
2. Feature Extraction: The ResNet backbone extracts feature maps from the input image.
3. Region Proposal: The RPN generates proposals for regions likely containing human figures.
4. RoI Extraction: The RoIAlign layer extracts aligned feature maps from these proposals.
5. Dense Correspondence Prediction: The DensePose head predicts dense correspondences (UV coordinates) for each pixel in the RoI feature maps.
6. UV Mapping: The UV coordinates are mapped to a 3D human model, resulting in dense pose estimations.

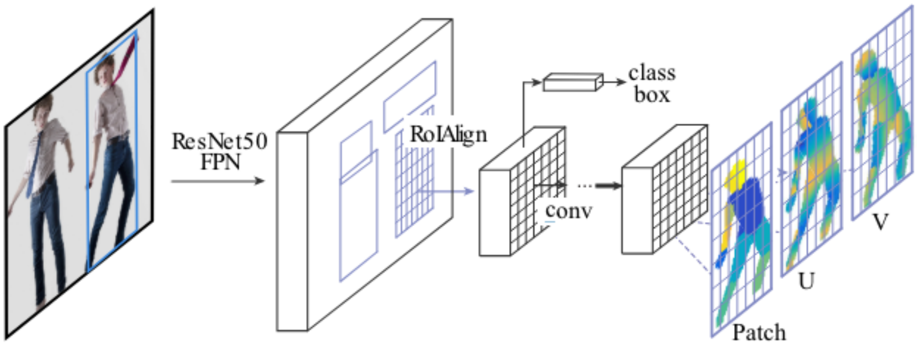

Figure 1: Structure of DensePose.

As shown in Figure 1, the fully-convolutional network introduced on top of RoI-pooling consists of 8 alternating layers. These layers are a combination of 3×3 fully convolutional layers and ReLU (Rectified Linear Unit) activation layers, each with 512 channels.

DensePose formulates the dense correspondence estimation as a combination of classification and regression tasks. The classification task assigns each pixel to either the background or one of the surface parts, while the regression task determines the exact coordinates of the pixel within the assigned part. The process is mathematically represented as:

$$c^* = \arg max_c \, P(c|i) \tag{3.1}$$

$$[U, V] = R_{c^*}(i) \tag{3.2}$$

where c^* is the body part with the highest posterior probability $P(c|i)$ for pixel i, and $R_{c^*}(i)$ provides the UV coordinates for the pixel within the part.

The classification branch employs a 25-way softmax classifier (24 parts + background), and the regression branch uses smooth L1 loss for coordinate prediction within each part. The regression loss is computed only for the pixels belonging to the respective parts. The classification loss ensures that each pixel is assigned to the correct body part, while the regression loss fine-tunes the exact position within that part.

DensePose enhances performance through cascading, where the output of an initial prediction stage is refined by subsequent stages. This iterative refinement provides contextual information and benefits from deep supervision. Multi-task learning is incorporated by sharing features between dense pose estimation, keypoint detection, and instance segmentation tasks, leveraging the complementary nature of these tasks. This multi-task learning approach allows the model to use shared representations, improving efficiency and accuracy across tasks.

The cross-cascading architecture integrates multiple tasks, where the initial predictions from DensePose, masks, and keypoints are combined and refined in subsequent stages. This approach ensures that each task benefits from the intermediate results of the others, leading to improved overall performance.

The training of DensePose involves several stages, starting with the annotation of dense correspondences in the DensePose-COCO dataset. Human annotators establish dense correspondences between 2D images and a 3D surface model, resulting in high-quality ground-truth data. The annotations are used to train the FCN, which is subsequently refined using a cascading approach. This approach ensures that the model learns from high-quality, dense annotations, improving its ability to generalize to new images.

During training, a teacher network is employed to inpaint the supervision signal on positions not originally annotated, creating a dense supervision signal. This process enhances the performance by providing a more comprehensive supervision signal, ensuring better generalization and accuracy. The teacher network fills in the gaps in the annotations, allowing the student network to learn from a more complete and dense dataset.

In the context of our project, DensePose plays a crucial role in accurately identifying the positions of human figures and their body parts within video frames. This is essential for the subsequent stages of our methodology, where gaze direction is determined to evaluate whether viewers are looking at advertisements. By leveraging DensePose, we can achieve precise localization of human body parts, particularly the head and eyes, which forms the foundation for accurate gaze estimation. This capability is crucial for confirming viewer engagement with advertisements.

3.2 Gaze360

Gaze360 is a sophisticated method designed for robust 3D gaze estimation in unconstrained environments [15]. The model and dataset are tailored to handle diverse real-world conditions, making it highly suitable for applications where traditional gaze estimation methods fall short. This section delves into the technical details of the Gaze360 methodology.

Gaze360 employs a combination of convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks to process sequences of video frames, providing accurate gaze direction estimates. The architecture of Gaze360 consists of several key components, each playing a crucial role in ensuring robust and accurate gaze predictions. The model begins with a ResNet-18 backbone, which is pretrained on ImageNet. This backbone processes input frames to extract high-level features. ResNet-18 is chosen for its balance between depth and computational efficiency, making it effective for feature extraction in dynamic scenes. The ResNet-18 architecture consists of 18 layers, including convolutional layers, batch normalization layers, ReLU activation layers, and pooling layers, all of which contribute to the extraction of detailed features necessary for accurate gaze estimation. Figure 2 illustrates the architecture of the Gaze360 model.

After feature extraction, the high-level features from multiple consecutive frames are fed into a bidirectional LSTM network. The LSTM network, consisting of two layers, models the temporal dependencies in the sequence of frames, allowing the model to account for the continuity and flow of gaze direction over time. Each LSTM layer contains 256 hidden units, which help in capturing both forward and backward temporal dependencies. This bidirectional approach ensures that the model can utilize information from both past and future frames to make more accurate predictions.

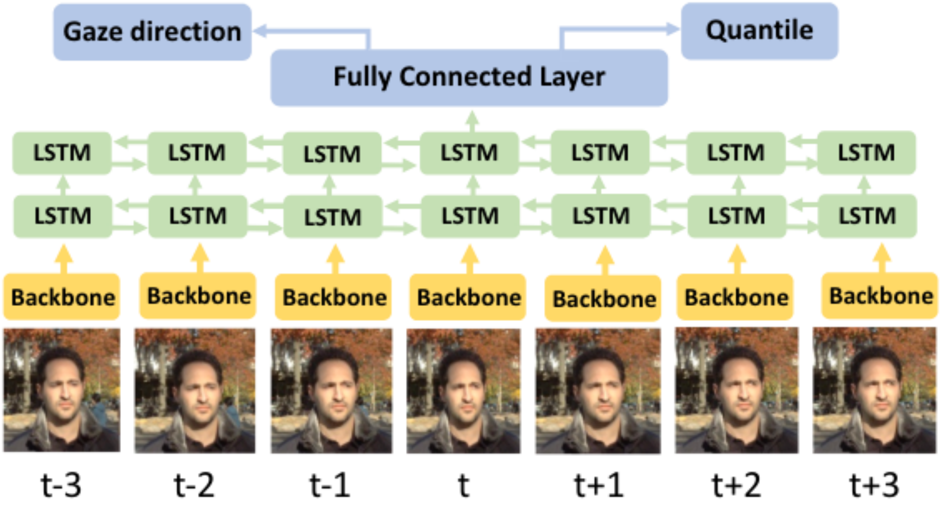

Figure 2: Gaze360 model architecture.

The outputs from the LSTM network are concatenated and passed through fully connected layers to produce the final predictions, including the gaze direction and an estimate of uncertainty. The fully connected layers are designed to map the high-level features and temporal information into meaningful gaze direction estimates. Specifically, the fully connected layers consist of 512 neurons, which further process the concatenated LSTM outputs before making the final predictions.

Gaze360 is trained on a comprehensive dataset comprising 238 subjects recorded in both indoor and outdoor environments. The dataset includes a wide range of head poses and distances, providing a diverse and comprehensive training set that enhances the model's generalization capabilities. The data collection setup involves a 360-degree camera system, allowing for simultaneous capture of multiple subjects from different angles. The setup includes a moving target board with an AprilTag marker to guide the subjects' gaze, ensuring accurate and varied gaze data collection. This comprehensive dataset ensures that the model can generalize well to different real-world scenarios.

The primary loss function used in training Gaze360 is the pinball loss, which is suitable for quantile regression and helps the model predict the expected gaze direction along with an uncertainty interval. The pinball loss is defined as follows:

$$\hat{q} = \begin{cases} \theta_{gt} - (\theta - a), & for\ \tau \leq 0.5 \\ \theta_{gt} - (\theta + a), & otherwise \end{cases} \tag{3.3}$$

$$L_T(\theta, \sigma, \theta_{gt}) = \max(\tau \hat{q}_\tau, -(1 - \tau)\hat{q}_\tau) \tag{3.4}$$

where:

- $\theta_{gt}\ and\ \phi_{gt}$ are the ground truth gaze angels,
- θ and ϕ are the predicted gaze angels,
- σ is the predicted uncertainty.

This formulation allows the model to learn both the gaze direction and the confidence intervals, which is critical for applications in unconstrained environments where prediction accuracy can vary. The model is trained using the Adam optimizer with a learning rate of 10^{-4} . The training process involves both supervised learning and self-supervised domain adaptation techniques to improve the model's performance on new, unseen data.

In our project, Gaze360 is used to determine whether viewers are looking at advertisements, thereby confirming their engagement. By accurately estimating the gaze direction, Gaze360 provides insights into viewer attention, enabling us to assess the effectiveness of advertising content in real-world scenarios. This application is crucial for understanding how advertisements capture viewer interest and ensuring that the content is engaging. Gaze360's robust architecture and comprehensive training enable it to provide reliable gaze predictions even in varied and unconstrained environments, making it a valuable tool for our advertising analysis.

3.3 ManyDepth

ManyDepth is an innovative self-supervised approach to monocular depth estimation that effectively utilizes multiple frames at test time to enhance depth prediction accuracy. This capability is critical for applications such as augmented reality, autonomous driving, and 3D reconstruction, where accurate depth information is paramount. In our project, ManyDepth plays a pivotal role in accurately estimating the spatial positions of objects within video frames, which is essential for determining whether viewers are looking at advertisements. By leveraging the sequence information available at test time, ManyDepth provides a robust solution for improving the accuracy of depth predictions, thus enhancing our ability to analyze viewer engagement with advertisements.

Developed by Watson et al., ManyDepth was introduced in their paper "The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth" [16]. The core innovation of ManyDepth lies in its ability to utilize sequence information at both training and test times. This approach is inspired by multi-view stereo techniques and incorporates a deep end-to-end cost volume-based method trained using self-supervision. The model adapts to the availability of sequence information at test time, providing superior depth predictions compared to single-frame methods.

The ManyDepth model architecture comprises several key components (Figure 3):

- Encoder: The encoder processes the input image to extract feature maps. The encoder is based on a ResNet architecture, known for its strong feature extraction capabilities.
- Decoder: The decoder reconstructs the depth map from the encoded feature maps.
- PoseCNN: The PoseCNN predicts the relative poses between frames, which is crucial for constructing the cost volume.
- Cost Volume Construction: The cost volume is built from the feature maps and relative poses, representing the depth hypotheses for each pixel.
- Loss Functions: The model is trained using a combination of photometric loss, smoothness loss, and a novel consistency loss designed to handle occlusions and moving objects.

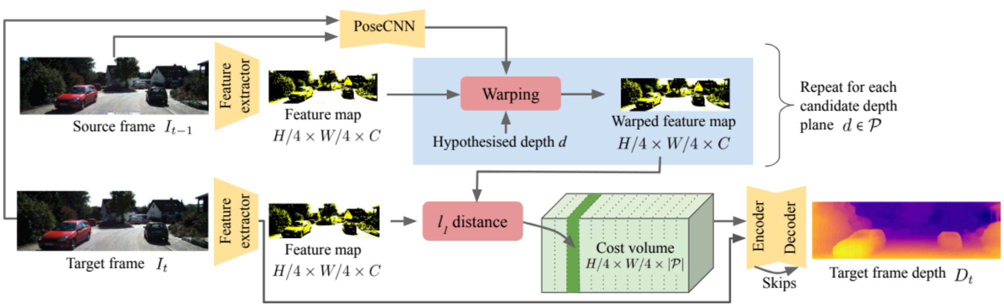

Figure 3: ManyDepth structure.

The ManyDepth model is trained using self-supervised learning, where the primary supervision signal comes from the reconstruction of adjacent frames. The training process involves the following key components:

- **Photometric Reprojection Loss:** This loss ensures the model correctly estimates depth by minimizing the reprojection error between the synthesized depth map and the actual depth map observed from multiple frames. It uses a minimum reprojection strategy which is more robust against occlusions and disocclusions:

$$L_p = min_{t'}(SSIM(I_t, I_{t'\to t}) + \lambda. \|I_t - I_{t'\to t}\|_1) \tag{3.5}$$

  - I_t is target image.
  - $I_{t'\to t}$ is is the synthesized view from frame $t'$to t, using the predicted depth and camera motion.
  - $SSIM$ is the structural similarity index, and λ is a parameter balancing $SSIM$ and L1 losses, typically set to balance structural and absolute differences.

- **Smoothness Loss:** Encourages the predicted depth map to be smooth spatially, especially within textureless regions, helping to prevent overfitting to image noise or artifacts:

$$L_{smooth} = |\nabla d_t|. e^{-|\nabla I_t|} \tag{3.6}$$

- d_t is the depth map at target frame t.
- ∇d_t and ∇I_t represent the image and depth gradients, with the exponential term reducing the influence of the loss in areas with high image gradients (likely edges).

- **Consistency Loss:** Introduced to ensure that predictions are consistent across different frames, improving robustness to dynamic objects and changes in the scene not related to the camera motion:

$$L_{consistency} = \sum_{t'} M_{t'} \|D_t - D_{t'}\| \tag{3.7}$$

- D_t and $D_{t'}$ are the depth predictions for the target frame and other frames respectively.
- $M_{t'}$ is a mask that identifies where the cost volume may be unreliable due to moving objects or occlusions.

The overall training loss is a weighted sum of these individual loss components:

$$L = (1 - M)L_p + L_{consistency} + L_{smooth} \quad (3.8)$$

In the context of our project, ManyDepth is used to accurately estimate the depth information in video frames, which is critical for determining whether viewers are looking at advertisements. By leveraging the sequence information available at test time, ManyDepth enhances the accuracy of depth predictions, enabling precise localization of objects and their spatial relationships within the scene. This capability is essential for analyzing viewer engagement with advertisements, as it allows us to accurately determine if viewers are looking at the advertisements or not.

3.4 Integrating Outputs in a 3D Environment for Decision Making

In this section, the integration process of the outputs from DensePose, Gaze360, and ManyDepth algorithms is described to create a comprehensive 3D projection. This 3D projection is essential for accurately determining whether viewers are looking at advertisements, thereby assessing the effectiveness of the advertisements.

The process begins by extracting the viewer’s eye coordinates using the DensePose algorithm, which provides precise localization of body parts in the 2D image. The midpoint of the viewer's eyes is used as the representative point for the viewer's gaze direction. The depth information for the viewer is derived from the ManyDepth algorithm. A bounding box encompassing the viewer’s head is identified using the DensePose output. This bounding box is then used to extract the corresponding depth values from the depth map generated by ManyDepth. The depth value for the viewer’s head is computed as the median of the depth values within the bounding box to avoid inaccuracies due to parts of the bounding box that might not belong to the head.

With the 2D coordinates and the corresponding depth value, the viewer's eye position is projected into the 3D space. The 3D coordinates are calculated using the intrinsic parameters of the camera and the depth information, providing a precise 3D location of the viewer's eye point in the scene.

Next, the advertisements in the scene are manually identified in the original 2D image. Bounding boxes are drawn around each advertisement. For each advertisement, the depth information is extracted using the ManyDepth model. The depth values at the corners of the bounding box are used to compute the 3D coordinates of the advertisement corners. These 3D coordinates are then used to place the advertisement accurately within the 3D projection.

The gaze direction of the viewer is predicted using the Gaze360 model. This model takes input from cropped images around the viewer’s head across several frames (t-3, t-2, t-1, t, t+1, t+2, t+3) to predict the gaze direction in spherical coordinates (θ, ϕ). These spherical coordinates are converted into Cartesian coordinates using the following mathematical transformation:

$$x = \cos(\phi) . \sin(\theta) \tag{3.9}$$

$$y = \sin(\phi) \tag{3.10}$$

$$z = -\cos(\phi) . \cos(\theta) \tag{3.11}$$

In this context:

- θ is the azimuthal angle.
- ϕ is the polar angle.
- x, y, z are the Cartesian coordinates.

The resulting 3D gaze vector extends from the viewer's eye point, indicating the direction of the gaze. The starting point of this vector is the viewer's eye point, and the vector extends outward, indicating the direction of the gaze.

The 3D coordinates of the viewer’s eye point, the 3D coordinates of the advertisements, and the gaze vector are integrated into a unified 3D projection. This projection is visualized using 3D rendering software. To determine whether the viewer’s gaze intersects with a specific advertisement, we use the following mathematical procedure:

First, the plane of the advertisement is defined by computing the edges of the advertisement's bounding box in 3D space:

$$edge_1 = corner_{br} - corner_{tl} \tag{3.12}$$

$$edge_2 = [0, 0, corner_{br_z} - corner_{tl_z} \tag{3.13}$$

The normal vector of the plane is calculated using the cross product of the edges:

$$n = edge_1 \times edge_2 \tag{3.14}$$

The coefficients A, B, and C are extracted from the normal vector, and D is calculated using the top-left corner of the bounding box:

$$D = -(n.corner_{tl}) \tag{3.15}$$

Next, the intersection of the gaze ray with the plane is calculated. The denominator is determined to check if the gaze direction is parallel to the plane:

$$denominator = A\,.gazedirection_x + B.gazedirection_y + C\,.gazedirection_z$$

If the denominator is not zero (i.e., the gaze is not parallel to the plane), the parameter t for the intersection point is computed:

$$t = -\frac{A\,.eyeposition_x + B.eyeposition_y + C.eyeposition_z + D}{denominator} \tag{3.16}$$

The intersection point is then calculated:

$$intersectionpoint = eyeposition + t\,.gazedirection \tag{3.17}$$

Finally, it is checked whether the intersection point lies within the bounds of the advertisement:

$$\begin{aligned} withinbounds & \\ &= \left(min\left(corner_{tl_x}, corner_{br_x}\right) \le intersectionpoint_x \right. \\ &\le \max\left(corner_{tl_x}, corner_{br_x}\right)) \wedge \left(min\left(corner_{tl_y}, corner_{br_y}\right)\right. \\ &\le intersectionpoin_y \le \max\left(corner_{tl_y}, corner_{br_y}\right)\Big) \\ &\wedge \left(min\left(corner_{tl_y}, corner_{br_y}\right) \le intersectionpoin_y \right. \\ &\le \max\left(corner_{tl_y}, corner_{br_y}\right)\Big) \end{aligned} \tag{3.18}$$

If the intersection point lies within these bounds, it indicates that the viewer is looking at the advertisement. This process allows for a detailed analysis of viewer engagement with advertisements, providing insights into which advertisements are effective in capturing attention.

By integrating the outputs from DensePose, Gaze360, and ManyDepth into a cohesive 3D model, we create a robust framework for accurately analyzing and understanding viewer behavior in relation to advertisements. This comprehensive approach enhances the understanding of spatial relationships and interactions within the scene, providing valuable data for optimizing advertising strategies based on precise viewer engagement metrics.

3.5 Data Acquisition for Testing

To effectively test the application's performance, the study employs a targeted approach to data acquisition. The primary sources of data are videos from public platforms such as YouTube and other relevant internet sources, specifically selected to include scenes that feature people interacting with billboards. These videos provide a realistic and diverse array of viewer behaviors and environmental conditions, which are critical for testing the application under varied real-world scenarios.

Additionally, to further enrich the dataset and tailor it to specific research needs, I have conducted personal video recordings. These recordings are designed to capture specific interactions between viewers and outdoor advertisements, focusing on varying distances, angles, and lighting conditions to thoroughly test the robustness and accuracy of the gaze detection, human pose estimation, and depth estimation technologies used in the study.

Given the public nature of the video content from online platforms, care is taken to ensure that only publicly available videos are used, where no privacy violations occur. For personal recordings, I have obtained the necessary consents where applicable, ensuring all data collection processes adhere to ethical standards. All data, particularly personally recorded footage, is handled with strict adherence to data protection regulations to ensure participant privacy and data security.

4. RESULTS: EXPERIMENTAL ANALYSIS

This section presents the results obtained from testing our integrated framework on a specific frame. The aim is to determine whether the viewer is looking at the advertisement by evaluating the outputs from DensePose, Gaze360, and ManyDepth models using our established algorithm.

Experimental Setup Description

During this test, the workflow described in the methodology section was implemented using Python programming language. The code was executed with the support of the National High-Performance Computing Center (UHeM) at Istanbul Technical University. These experiments were conducted on a specific node with an Intel XEON 8362 processor, 512 GB of RAM, and an NVIDIA A100 80 GB GPU, enabling parallel processing.

The test involved a video recorded on my phone, featuring a person and an advertisement. Each frame of the video was processed through each algorithm. The overall processing time is summarized in the table below:

Model Name	Frame Count	Total Elapsed Time (ms)	E.T. perFrame (ms)
DensePose	250	30148,11	120,59
Gaze360	250	14035,83	20,25
ManyDepth	250	19925,10	80,02
Decision Algorithm	250	255,28	0,36

Based on these results, the total time to process a single frame is 120.5924 + 20.2537 + 80.0205 + 0.3684, which equals 221.235 ms. This duration is approximately one-fifth of a second, sufficient for real-time processing. The most time-consuming algorithms, DensePose and ManyDepth, can be optimized by reducing the number of layers and neurons.

Presentation of Experimental Data

An example from the video is presented step by step below. First, we start with the image in its original dimensions, 1920×1080. Original image is show in Figure 4. This image was initially processed with the DensePose algorithm, which detected the person and identified the bounding box around the head and the eye coordinates. The bounding box is shown in Figure 5.

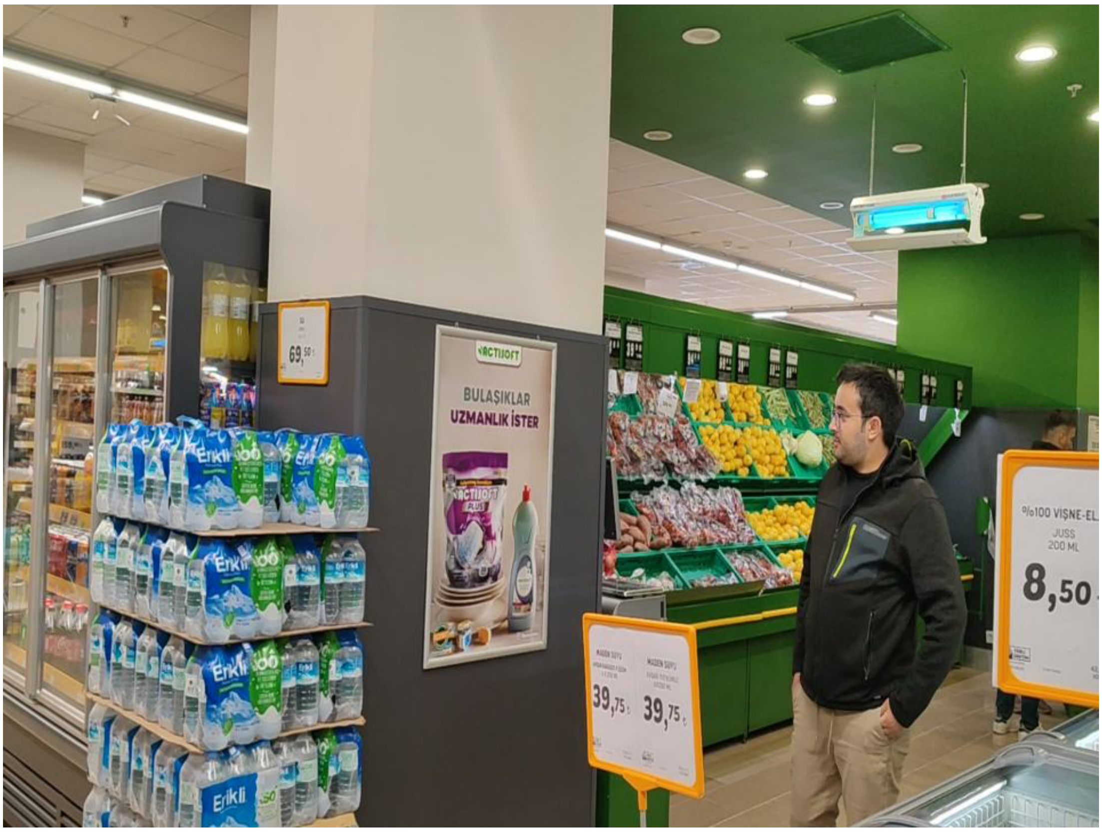

Figure 4: Original image

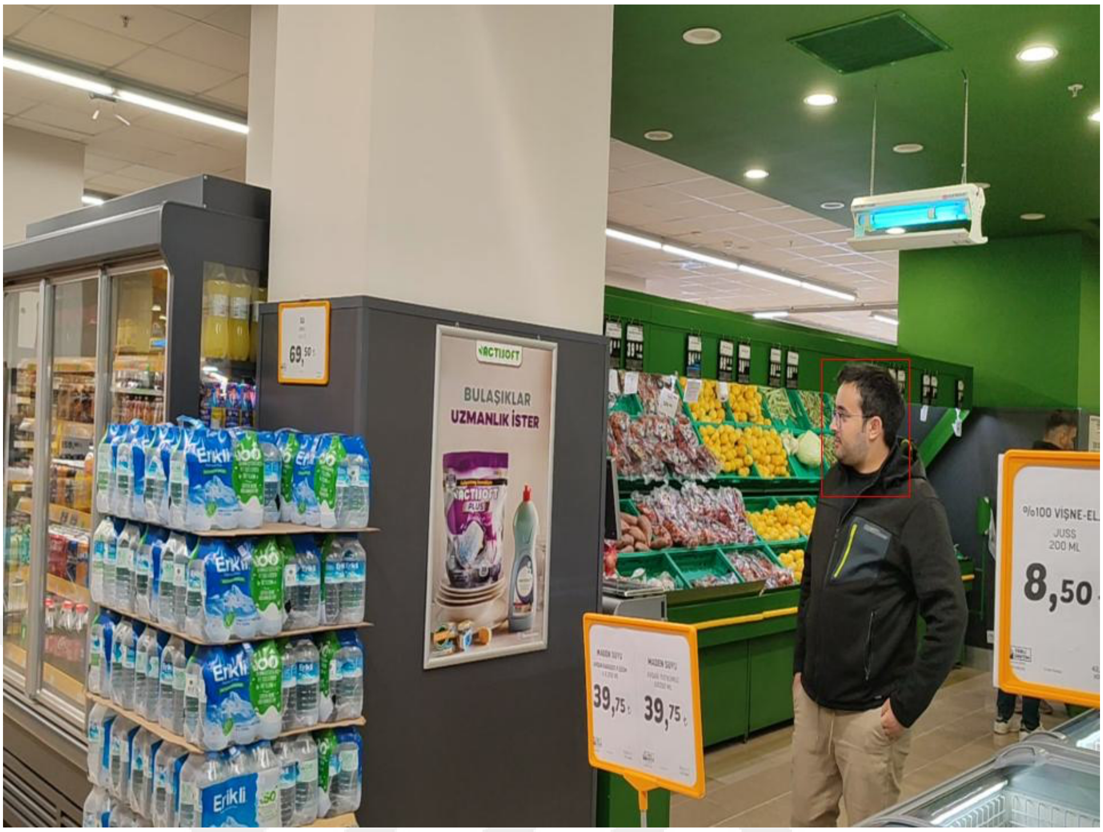

Figure 5: Bounding box obtained from DensePose output.

To be suitable for the Gaze360 model and to speed up processing, the image dimensions were rescaled to 960 x 720. Subsequently, the bounding box was cropped from the image to obtain a smaller image containing only the person. Additional frames at t-3, t-2, t-1, and t+1, t+2, t+3 were also obtained. These frames are shown in Figure 6. For real-time prediction, if t+1, t+2, and t+3 frames are unavailable, they will be filled with the frame at time t. This input image was then fed into the Gaze360 model, which predicts the gaze direction in spherical coordinates. These spherical coordinates are converted to Cartesian coordinates to obtain the gaze direction. The gaze direction vector is shown in Figure 7.

Figure 6: Input images provided to the Gaze360 model.

Figure 7: Gaze direction vector.

Next, the image is processed with the ManyDepth model to generate the depth map. The bounding box and eye location are matched in this depth map to find the corresponding depth values. Using the eye location as the starting point, the endpoint of the gaze direction vector is calculated using the formula: start_point + arrow_length * gaze_direction. The arrow length is typically configured based on the camera setup, which is 250 in this example. This allows the arrow to be drawn to the end of the image. After determining the endpoint, all data can be projected into a 3D space. This 3D projection is shown in Figure 9. The final location of the advertisement is manually annotated, as seen in Figure 8. Assuming the camera is stationary, this step only needs to be performed once according to the camera settings. With all this data, the decision algorithm is executed to determine if the person is looking at the advertisement.

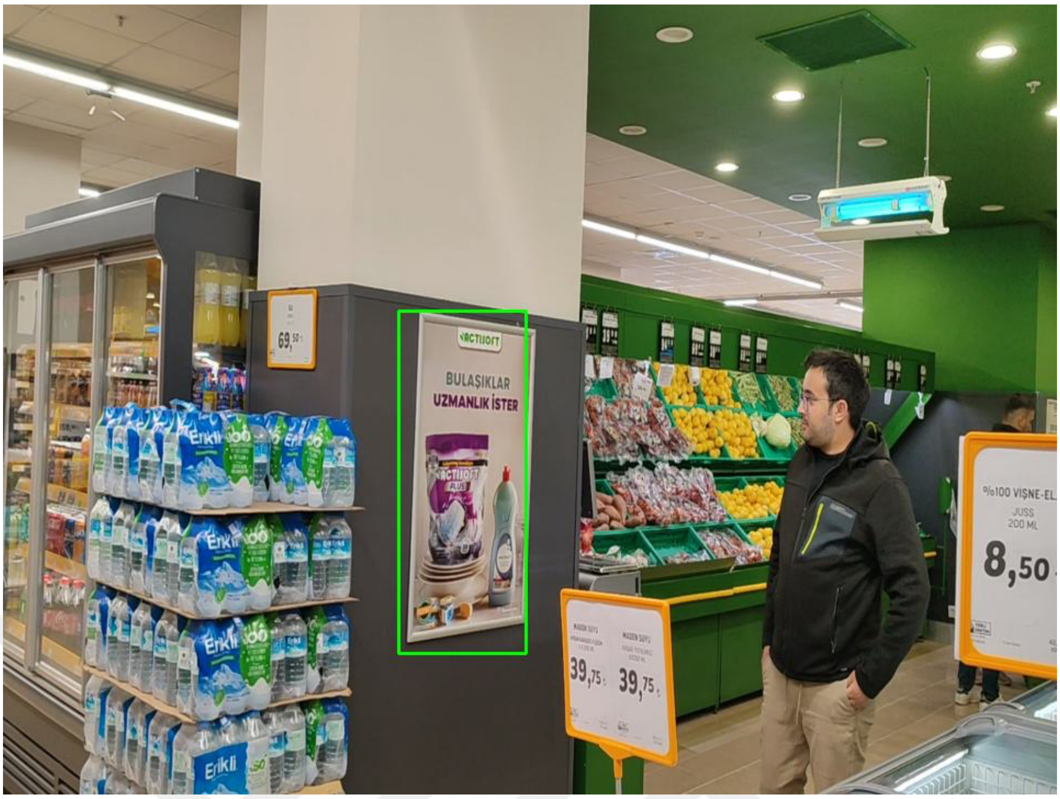

Figure 8: Manually annotated advertisement location.

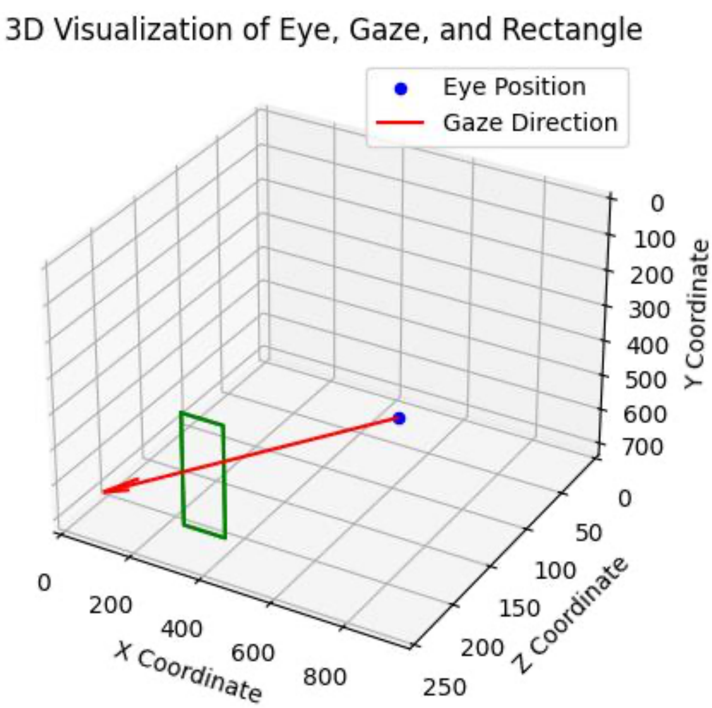

Figure 9: 3D projection.

These results demonstrate the effectiveness of our integrated framework in determining whether the viewer is looking at the advertisement. Future work may focus on improving the depth estimation process and testing this framework on a larger dataset to validate these findings further.

5. CONCLUSION

This thesis has focused on developing an innovative framework that integrates DensePose, Gaze360, and ManyDepth models to analyze viewer engagement with outdoor advertisements. The primary aim was to create a robust system that accurately determines whether viewers are looking at advertisements based on their gaze direction and the spatial context of the scene.

Our experimental results have shown that this integrated framework is highly effective in predicting viewer attention. By processing video frames to extract human poses, gaze directions, and depth information, we were able to construct a detailed 3D projection of each scene. This 3D projection enabled precise analysis of the intersection between the viewer's gaze vector and the advertisement's spatial position. The findings indicate that:

- The combination of DensePose, Gaze360, and ManyDepth provides a comprehensive approach to understanding viewer engagement.
- The system's real-time processing capabilities are sufficient for immediate feedback, which is crucial for dynamic advertising scenarios.
- The intersection analysis method effectively quantifies viewer engagement, offering valuable metrics for advertisers to evaluate and optimize their campaigns.

This work makes significant contributions to the fields of computer vision and digital advertising by:

1. Demonstrating the feasibility of integrating multiple advanced vision models to enhance the analysis of human attention in real-world scenarios.
2. Providing a methodological framework that can be applied to other domains requiring analysis of human interaction and engagement.

3. Offering insights into optimizing computational resources for real-time processing in applied deep learning projects.

While the results are promising, several limitations were identified that could be addressed in future work:

- **Model Optimization:** The DensePose and ManyDepth models, being the most time-consuming components, offer opportunities for optimization. Reducing their computational complexity while maintaining accuracy is a key challenge.
- **Dataset Variability:** Testing the framework on a more diverse set of video data would help assess its effectiveness across different environments and viewer demographics.
- **Algorithm Generalization:** Further research is needed to generalize the algorithm to handle various advertisement formats and placements, enhancing its applicability in the advertising industry.

In summary, this thesis has laid the groundwork for using integrated computer vision techniques to measure viewer engagement with advertisements. It opens up new possibilities for advertisers to tailor their content based on real-time engagement metrics, leading to more effective advertising strategies. Future research could significantly influence both academic study and practical applications in digital marketing and beyond.

REFERENCES

[1] **Statista.** (2024). Global advertising market - statistics & facts. Retrieved from https://www.statista.com/topics/990/global-advertising-market/#topicOverview

[2] **Statista.** (2024). Online advertising spending worldwide. Retrieved from https://www.statista.com/statistics/237974/online-advertising-spending-worldwide/

[3] **75media.** (2024). How many ads do we see in one day? Retrieved from https://75media.co.uk/blog/how-many-ads-seen-one-day/

[4] **Efeturk, O. A., Turgut, G., Asadi Dereshgi, H., & Yılmaz, A.** (2022). A Review of Visual Attention Research Using Eye-Tracking Technologies.

[5] **Wu, T.** (2016). The Attention Merchants: The Epic Scramble to Get Inside Our Heads.

[6] **Falkinger, J.** (2008). Limited attention as a scarce resource in information-rich economies.

[7] **Chun, M. M., & Turk-Browne, N. B.** (2007). Interactions between attention and memory.

[8] **Rayner, K.** (1998). Eye movements in reading and information processing: 20 years of research.

[9] **Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P.** (2004). Image quality assessment: from error visibility to structural similarity.

[10] **Sitzmann, T., & Weinhardt, J.** (2014). Training Engagement Theory: A Multilevel, Temporal Perspective on the Effectiveness of Training.

[11] **Wedel, M., & Pieters, R.** (2008). Eye Tracking for Visual Marketing.

[12] **Rayner, K.** (2009). Eye movements and attention in reading, scene perception, and visual search.

[13] **Pieters, R., Wedel, M., & Batra, R.** (2010). The Stopping Power of Advertising: Measures and Effects of Visual Complexity.

[14] **Guler, R. A., Neverova, N., & Kokkinos, I.** (2018). DensePose: Dense Human Pose Estimation In The Wild.

[15] **Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A.** (2019). Gaze360: Physically Unconstrained Gaze Estimation in the Wild.

[16] **Watson, J., Mac Aodha, O., & Brostow, G. J.** (2021). Temporal Consistency Loss for High Resolution Monocular Depth Estimation.

[17] **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning.

CURRICULUM VITAE

Name Surname : Fatih Baday

EDUCATION :

- **B.Sc.** : 2019, Gaziantep University, Faculty of Engineering, Department of Industrial Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2019-2021 Worked as a Data Scientist at Çalık Holding.
- 2021-2023 Worked as a Python Developer at Huawei.
- 2023-2024 Worked as a Machine Learning Engineer at Hangikredi.