DISSERTATION

VISION BASED ARTIFICIAL INTELLIGENCE FOR OPTIMIZING E-COMMERCE EXPERIENCES IN VIRTUAL REALITY

Submitted by

Panteha Alipour

Department of Systems Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2025

Doctoral Committee:

Advisor: Erika Gallegos

Thomas Bradley
Marie Vans
Mohammed Arefin

ABSTRACT

VISION BASED ARTIFICIAL INTELLIGENCE FOR OPTIMIZING E-COMMERCE EXPERIENCES IN VIRTUAL REALITY

Advancements in artificial intelligence (AI) and digital technologies have deeply reshaped consumer behavior and marketing strategies, demanding innovative approaches to decoding and optimizing customer engagement. This dissertation explores the potential of vision deep neural networks, generative AI, and virtual reality (VR) to analyze emotional and behavioral responses and enhance strategic business insights in digital commerce.

This research focuses on convolutional neural network (CNN) architectures and evaluates their effectiveness in predicting consumer engagement through facial emotion recognition (FER). The dissertation addresses limitations in FER datasets by integrating synthetic data generated using generative adversarial networks (GANs) and real-world open data extracted from social media. This hybrid approach enhances model generalizability across diverse demographics and advertisement categories.

The dissertation further investigates the role of immersive VR environments in influencing consumer engagement and purchase intent. By leveraging multi-modal causal analysis, it examines the interplay between VR design complexity, exposure sequencing, and emotional responses, providing actionable insights for optimizing e-commerce experiences.

Ethical considerations are central to this research, which address biases, privacy concerns, and transparency in AI-driven decision-making. The findings contribute to the development of robust, inclusive, and scalable frameworks for personalized commerce, offering a transformative approach to understanding consumer behavior in digital environments.

Through a systematic integration of vision deep learning, generative AI, and VR technologies, this dissertation bridges critical gaps in systems engineering research and business applications;

advancing both theoretical understandings and practical applications in consumer engagement optimization.

ACKNOWLEDGEMENTS

A special thanks to my advisor, Dr. Erika E. Gallegos, for being a source of guidance and inspiration throughout this entire journey. Dr Gallegos' intuitional feedback and thoughtful direction have been pivotal in shaping the trajectory of this dissertation.

Dr. Gallegos' commitment to excellence coupled with her ability to inspire has strengthened my research. I am incredibly grateful for her mentorship, and this dissertation wouldn't be what it is without her guidance.

DEDICATION

To the infinite game of learning, may curiosity forever guide the way.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
DEDICATION v
LIST OF TABLES ix
LIST OF FIGURES x

Chapter 1 Introduction 1
1.1 Research Aims and Research Questions 2
1.1.1 Research Aim 1 Description 2
1.1.2 Research Aim 2 Description 3
1.1.3 Research Aim 3 Description 3
1.1.4 Research Aim 4 Description 4
1.2 Contributions to Knowledge 4

Chapter 2 Background 6
2.1 Introduction to AI in Business Decision Making 6
2.1.1 AI Powered Marketing Innovations 7
2.2 Facial Emotion Recognition (FER) in Consumer Behavior Analysis 8
2.2.1 Facial Emotion Recognition (FER) in Marketing Research 10
2.2.2 Challenges in Facial Emotion Recognition 11
2.3 Data Augmentation and Generalizability in Vision Deep Learning 11
2.4 Generative AI for Dataset Expansion in Vision Deep Learning 13
2.5 Virtual Reality (VR) and E-Commerce Business 15
2.6 Challenges in AI-Powered E-Commerce Research 17
2.7 Gaps in Literature 18

Chapter 3 Data Overview 20

Chapter 4 Aim 1: Effectiveness of CNNs in Predicting Consumer Engagement 22
4.1 Aim 1 Summary 22
4.2 Introduction 22
4.2.1 Problem Statement 24
4.3 Analytical Methodology 24
4.3.1 Convolutional Neural Network Architecture 25
4.3.2 Basic Convolutions Neural Network Components 27
4.4 Data Overview 32
4.4.1 NeuroBioSense Dataset 32
4.4.2 Dataset Used for Training the Facial Emotion Recognition Model 34
4.4.3 Data Pre-Processing 35
4.4.4 Fine Tuning and Data Augmentation 37
4.5 Results and Discussion 38

4.5.1 Convergence of Binary Classifier Over Training Epochs . . . 38
4.5.2 Evaluation Metrics . . . 41
4.5.3 Evaluation of Misclassified Instances . . . 46
4.5.4 Correlation of Emotions with Interest Levels . . . 47
4.6 Conclusions . . . 50

Chapter 5 Aim 2: Addressing Dataset Limitations with Generative AI and Social Media Data . . . 53
5.1 Aim 2 Summary . . . 53
5.2 Introduction . . . 53
5.2.1 Problem Statement . . . 54
5.3 Data Description . . . 55
5.3.1 NeuroBioSense (Baseline) Dataset . . . 55
5.3.2 NeuroBioSense and YouTube Combined (Baseline + YouTube) Dataset . 57
5.3.3 NeuroBioSense and Generative AI-Enhanced Synthetic Dataset (Baseline + StyleGAN2) . . . 58
5.3.4 Dataset Combinations for Models . . . 58
5.4 Theoretical Framework: StyleGAN2 . . . 60
5.5 Model Training and Optimization Strategy . . . 64
5.5.1 Data Augmentation and Regularization . . . 65
5.6 Experimental Design . . . 65
5.7 FER Model Framework . . . 66
5.7.1 Xception . . . 66
5.7.2 Xception Model Architecture . . . 67
5.7.3 Implementation Details . . . 68
5.7.4 Metrics for Evaluating Model Performance . . . 68
5.8 Results and Discussion . . . 70
5.8.1 Loss . . . 70
5.8.2 Accuracy . . . 71
5.8.3 Precision-Recall and Receiver Operating Characteristic Curve . . . 73
5.9 Ethical Considerations and Solutions . . . 75
5.9.1 Biases in Facial Emotion Recognition (FER) Models . . . 76
5.9.2 Privacy Concerns in Social Media Data . . . 76
5.9.3 Privacy Concerns in Synthetic Data . . . 77
5.9.4 Addressing Ethical Issues and Promoting Fairness . . . 77
5.9.5 Future Ethical Directions in FER Research . . . 78
5.10 Conclusions . . . 79

Chapter 6 Aim 3: Effects of VR Design Complexity and Exposure Sequencing on Engagement . . . 80
6.1 Aim 3 Summary . . . 80
6.2 Introduction . . . 80
6.2.1 Problem Statement . . . 81
6.3 Methodology . . . 82
6.3.1 Participants . . . 82

6.3.2 VR Environments . . . 82
6.3.3 Participant Groups . . . 84
6.3.4 Data Collection Procedure . . . 84
6.3.5 Data Quality and Integrity . . . 85
6.4 Results . . . 85
6.4.1 Summary of Pre-Exposure Variables . . . 85
6.4.2 Effects of Individual Differences on Perceived Engagement . . . 88
6.4.3 Effects of VR Environment Design on Perceptions . . . 92
6.4.4 Effects of VR Perceptions on Likelihood to Purchase . . . 94
6.4.5 Effects of Ordering on Perceived Engagement . . . 95
6.5 Discussion . . . 101
6.6 Conclusions . . . 103

Chapter 7 Aim 4: Extending Vision Deep Learning Models to VR . . . 106
7.1 Aim 4 Summary . . . 106
7.2 Introduction . . . 106
7.2.1 Problem Statement . . . 108
7.3 Data Collection and Description . . . 109
7.3.1 Procedures . . . 109
7.3.2 Participants . . . 110
7.3.3 Video Data . . . 110
7.4 Analytical Methodology . . . 111
7.5 Results and Discussion . . . 114
7.5.1 Receiver Operating Characteristic (ROC) Curves for Each Method . . . 114
7.5.2 Overall Performance for Each Method . . . 116
7.6 Conclusions . . . 117

Chapter 8 Conclusions . . . 119
8.1 Research Contributions . . . 119
8.2 Revisiting Research Aims and Research Questions . . . 120
8.2.1 Aim 1 Findings . . . 120
8.2.2 Aim 2 Findings . . . 121
8.2.3 Aim 3 Findings . . . 121
8.2.4 Aim 4 Findings . . . 122
8.3 Limitations . . . 123
8.4 Future Work . . . 124
8.5 Publications of Results . . . 126

Bibliography . . . 128

LIST OF TABLES

3.1 Data Used for Each Research Aim 20

4.1 Demographics and Advertisement Categories in the NeuroBioSense Dataset 33
4.2 Summary of Interested and Not-Interested Videos by Ad Category 33
4.3 Xception Model Confusion Matrix 44
4.4 ResNet-50 Model Confusion Matrix 44
4.5 Xception Model Classification Report 45
4.6 ResNet50 Model Classification Report 45
4.7 T-Test Results Comparing Mean Emotion Probabilities for Interested and Not Interested Classes 50

5.1 Number of Images Used in Training, Validating, and Testing Each Model 59
5.2 Performance Improvements for Precision-Recall AUC and ROC-AUC Across Models . 75

6.1 Frequency Distribution of Gender Across Age Groups 82
6.2 Summary of Linear Mixed Model Predicting Perceived Engagement 89
6.3 Summary of Linear Mixed Model Predicting Likelihood to Purchase 95
6.4 Summary of ASMD Values Before and After Matching, and Reduction Percentages for Each Covariate 97

7.1 Participant Ages and Genders 110
7.2 Summary Statistics of the Dataset 111
7.3 Frame and Temporal Level Classification Methods Employed 112
7.4 Overall Performance Comparison of the Different Aggregation Methods 117

LIST OF FIGURES

4.1 ResNet-50 layer structure. 26
4.2 Xception layer structure. 27
4.3 A proposed CNN architecture for engagement detection classification. 28
4.4 Distribution of participant ages based on interest levels across ad categories. 34
4.5 Data pre-processing pipeline used for analyzing video reactions. 36
4.6 Distribution of data split for training, validation, and testing. 36
4.7 Loss curve for training (blue) and validation (orange) data for Xception (left) and ResNet-50 (right) models. 39
4.8 Accuracy curve for training (blue) and validation (orange) data for Xception (left) and ResNet-50 (right) models. 40
4.9 Precision-recall curves for the different models. 42
4.10 Receiver Operating Characteristic (ROC) curves for the different models. 43
4.11 Analysis of Xception model inaccuracies in detecting emotional engagement. 46
4.12 Pairwise distribution and correlation of emotional responses between interested and not-interested classes. 48
4.13 Density plots of emotional response probabilities for interested and not interested participants. 49

5.1 Overview of the training and validating pipeline integrating NeuroBioSense (Baseline), Social Media (YouTube), and Generative AI (StyleGAN2) datasets. 60
5.2 Synthetic image generation using a StyleGAN model based on 'interested' and 'not interested' image categories. 63
5.3 Comparison of training and validation losses across the three datasets over 300 epochs. 71
5.4 Comparison of training and validation accuracies across three datasets over 300 epochs. 72
5.5 PR (left) and ROC (right) curves illustrating models' performance and balance across different datasets. 74

6.1 Example of the two VR environments for the Tesla products. 83
6.2 Example products in VR Detailed for the two brands. 83
6.3 Correlation heatmap of pre-exposure variables. 86
6.4 Pair plot of pre-exposure variables by gender. 88
6.5 Partial dependence plots (PDPs) for predictors influencing perceived engagement in VR Detailed. 91
6.6 Partial dependence plots (PDPs) for predictors influencing perceived engagement in VR Simple. 92
6.7 Density plots comparing post-exposure perceptions between VR Simple and VR Detailed. 93
6.8 Density plot of VR Detailed engagement scores by treatment group for propensity matching scores. 99
6.9 Empirical cumulative distribution functions of VR Detailed engagement scores by treatment group. 100

7.1 Experiment workflow for vision deep learning comparisons of user engagement in VR. 110
7.2 ROC curves comparing the six methods for aggregating frame-level probabilities. . . . 115

Chapter 1

Introduction

Artificial intelligence (AI) and virtual reality (VR) are transforming how businesses engage with consumers. As business strategies evolve, the ability to analyze consumer behavior through data-driven insights becomes increasingly essential. This dissertation explores how vision-based deep learning and VR technologies can advance our understanding and optimization of consumer interactions.

The primary motivation for this research lies in addressing critical challenges in consumer behavior analysis. Traditional methods of understanding consumer responses, such as surveys or focus groups, are often limited by scalability, bias, and lack of granularity. Emerging AI technologies such as vision deep learning, and even more specifically convolutional neural networks (CNNs), generative adversarial networks (GANs), and immersive VR environments, provide transformative opportunities to overcome these challenges.

Understanding consumer interest and engagement remains a critical challenge in digital commerce. While facial expressions provide valuable nonverbal cues for decoding consumer behaviors, accurately interpreting these signals using AI is an underexplored frontier. Traditional datasets and models often lack the robustness needed to handle real-world scenarios, limiting their practical utility in dynamic and diverse markets.

This dissertation develops methodologies that integrate vision-based deep learning models, generative adversarial networks, and VR environments to decode emotional and behavioral responses, ultimately enhancing business strategies.

By addressing gaps in the literature, this research contributes to a deeper understanding of:

- The role of vision deep learning models in predicting consumer engagement.
- The potential of generative AI and social media data to expand datasets and improve model generalizability.

- The specific emotions extracted from facial expressions that correlate strongly with consumer interest and disinterest in advertisements.
- The impacts of design complexity in VR environments on consumer behavior.
- The integration of vision-based deep learning, generative AI, and VR environments into robust, scalable frameworks for personalized marketing.

1.1 Research Aims and Research Questions

This dissertation is structured around four primary aims, each linked to specific research questions (RQs). Together, these research questions seek to improve the use of AI and VR for consumer and e-commerce interactions.

1.1.1 Research Aim 1 Description

The objective of Research Aim 1 is to investigate the effectiveness of convolutional neural networks (CNNs), using the two prominent architectures of ResNet-50 and Xception, in analyzing facial expressions to predict consumer engagement during advertisement viewing. The results of this research aim have been published in the *International Journal of Human-Computer Interaction*, article titled "AI-Driven Marketing Personalization: Deploying Convolutional Neural Networks to Decode Consumer Behavior." The following research questions are addressed in Research Aim 1:

- **RQ 1.1:** How effective are two prominent convolutional neural network (CNN) architectures, Xception and ResNet-50, in distinguishing consumer engagement (interested vs. disinterested)?
- **RQ 1.2:** What role do specific emotional cues (e.g., happiness, disgust, fear, anger, etc.) play in consumer interest classification?
- **RQ 1.3:** What are the implications of facial expression analysis findings for personalized marketing strategies in digital advertising?

1.1.2 Research Aim 2 Description

The objective of Research Aim 2 is to assess the potential of generative AI and social media data to address dataset limitations and enhance the generalizability of vision deep learning models. The results of this aim have been published in the *Journal of Artificial Intelligence Review*, article titled "Leveraging Generative AI Synthetic and Social Media Data for Content Generalizability to Overcome Data Constraints in Vision Deep Learning." The following research questions are addressed in Research Aim 2:

- **RQ 2.1:** Can FER model generalizability be improved using data extracted from social media (YouTube) and/or generated using AI (GANs), as compared to controlled data from a laboratory study?
- **RQ 2.2:** How can FER models trained on specific categories of advertisements be generalizable to new categories of advertisements?
- **RQ 2.3:** What are the ethical and practical considerations in using synthetic and real-world data for FER model training?

1.1.3 Research Aim 3 Description

The objective of Research Aim 3 is to explore the effects of VR design complexity and exposure sequencing on consumer engagement and purchase intent. The results of this paper have been submitted to the *Journal of Business Research*, paper titled "A Multi-Modal Causal Analysis of Product Exploration: Examining the Impact of Immersion and Exposure Sequences on Consumer Behavior in Virtual Reality E-Commerce." The following research questions are addressed in Research Aim 3:

- **RQ 3.1:** How do individual differences, such as technological savviness, brand familiarity, and frequency of online gaming and shopping, moderate the relationship between VR complexity and engagement?

- **RQ 3.2:** How does designed visual complexity and opportunities for engagement within VR affect perceived immersion, engagement, realism, sense of presence, distraction, effort, and purchase intent?

- **RQ 3.3:** How do perceptions of VR interactions, such as perceived immersion, engagement, realism, etc., influence likelihood to purchase?

- **RQ 3.4:** What are the cognitive and emotional effects of exposure sequencing from a simpler VR to a more detailed VR environment and vice-versa?

1.1.4 Research Aim 4 Description

The objective of Research Aim 4 is to extend vision deep learning models into immersive virtual reality environments to analyze emotional responses and engagement while interacting with VR e-commerce. The results of this research aim plan to be submitted to the journal of *Virtual Reality*, article titled "AI-powered virtual reality: Enhancing user experience in VR e-commerce through facial emotion recognition." The following research question is addressed in Research Aim 4:

- **RQ 4.1:** How does accuracy compare for frame-level versus temporal-level CNN-based FER models in classifying user interest/disinterest in interactive VR environments?

1.2 Contributions to Knowledge

This dissertation makes significant contributions to the field of artificial intelligence (AI) in e-commerce using a systems engineering approach, particularly in understanding and predicting consumer engagement through Facial Emotion Recognition (FER). It addresses key gaps in the literature by advancing the generalizability of deep learning models and introducing novel methodologies for integrating diverse datasets. Specifically, this research establishes how the combination of controlled laboratory data, reaction videos sourced from social media, and synthetic data generated through Generative Adversarial Networks (GANs) can create more robust FER systems.

By leveraging these complementary data sources, the research overcomes biases and limitations associated with traditional datasets, which enables models to perform reliably across diverse demographic and contextual scenarios.

A central innovation of this research lies in its hybrid data strategy, in which synthetic data is used to enrich underrepresented emotional categories. Through the application of advanced generative techniques such as StyleGAN2, this approach ensures demographic and contextual diversity in the training process, addressing long-standing challenges of data set scarcity and lack of inclusion. The integration of real-world data with synthetic enhancements also provides a scalable and ethical solution to advance FER applications in business and engineering.

In deploying convolutional neural network architectures, including Xception and ResNet-50, this dissertation deepens the understanding of the relationship between specific emotional expressions and consumer interest. By identifying emotions such as happiness as strong predictors of engagement, and disgust as an indicator of disinterest, the findings offer actionable insights into the emotional dynamics of consumer decision-making. These results not only validate the efficacy of FER for marketing purposes but also provide a foundation for developing more personalized and effective advertising strategies.

This research also advances the conversation about the generalizability of facial expression recognition models across different advertising categories. By showing that models trained on specific datasets can accurately predict engagement with new and unseen content, the dissertation provides real-world evidence of these systems' adaptability in dynamic and diverse business environments. This adaptability is especially important for deploying AI in settings where content changes frequently and unpredictably.

In essence, this dissertation bridges the gap between cutting-edge developments in deep learning and their real-world applications in understanding consumer behavior. It makes a meaningful contribution to the growing field of AI-driven e-commerce by offering a practical framework for enhancing user engagement and stronger connections between consumers and brands.

Chapter 2

Background

2.1 Introduction to AI in Business Decision Making

The rapid digitization of commerce has transformed the way businesses operate, pushing them to rethink their strategies [1]. More than ever, companies are leaning on artificial intelligence (AI) to boost customer engagement through smarter and data-driven insights [2]. Due to the fast advancement of AI technologies, businesses can now analyze and make inferences from huge amounts of consumer data and identify patterns in ways that were previously unattainable [2]. Among the most impactful AI-driven techniques are deep learning methodologies, especially convolutional neural networks (CNNs), which have demonstrated their excellence in analyzing and interpreting complex data such as images and videos [3].

CNNs can revolutionize the field of business analytics by allowing businesses to decode consumer behavior [4]. By using CNNs to analyze facial expressions while people watch advertisements, companies can gain a clearer understanding of their audience's emotions and engagement levels [4]. These insights are critical in today's competitive digital economy, where consumer attention spans are limited and personalized content is a key differentiator [5]. By leveraging CNNs, marketers can assess real-time emotional cues and preferences; facilitating targeted campaigns that resonate deeply with specific audience segments [6].

AI extends beyond consumer engagement and can support real-time decision-making processes. Predictive analytics, powered by machine learning algorithms, enables businesses to anticipate consumer needs and dynamically adapt their strategies [7]. For example, recommendation systems on platforms like Amazon and Netflix utilize AI to analyze user behavior and preferences to deliver personalized suggestions that drive sales and customer satisfaction [8]. In addition, AI algorithms are employed to analyze social media trends, identify emerging consumer demands, and forecast market behavior, thereby giving businesses a competitive edge [9].

AI is also transforming how businesses operate by automating workflows and simplifying everyday tasks. AI has been used to streamline essential operations like content creation and email personalization [10]. For instance, tools powered by natural language processing (NLP) can generate personalized product descriptions, while reinforcement learning algorithms help optimize ad placements and budgets [11]. These innovations decrease operational costs and boost the efficiency of marketing campaigns, giving businesses more room to focus on strategic initiatives that drive long-term growth. [12].

Despite its transformative potential, the adoption of AI in businesses comes with its own set of challenges and ethical concerns. Issues like data privacy and algorithmic bias are drawing increasing scrutiny [13]. As companies collect and analyze consumer data, it has become more important than ever to ensure transparency and gain the trust of individuals by seeking their consent for AI applications [14]. At the same time, while AI brings undeniable benefits in terms of scalability and accuracy, its implementation is not without hurdles. Businesses must navigate the significant costs associated with technology upgrades and infrastructure investments [15].

2.1.1 AI Powered Marketing Innovations

The integration of AI into marketing has transformed how brands interact with consumers [1]. For example, Liu et al. [3] focused on quantifying how brands are portrayed on Instagram by analyzing consumer-generated images with classifiers trained on a Flickr dataset. This approach allows for the categorization of images based on specific brand attributes, thus enabling an understanding of how consumers visually represent brands in their social media posts. Similarly, Li et al. [1] sought to standardize the quantification of visual information by examining the effects of visual variation and content, which yielded a foundational methodological framework for future research on video marketing effectiveness.

Facial emotion recognition technologies leverage deep neural networks to capture and analyze customer reactions, providing real-time feedback on their engagement and satisfaction. This technology allows for more personalized marketing interactions and enhanced customer service by

adjusting the approach based on the consumer's emotional response [16]. Furthermore, AI's ability to process and analyze large volumes of data from various sources like text, images, and video content has proven that this capability supports more accurate and comprehensive market analyses and consumer insight gathering [17]. Thus, facilitating more effective marketing strategies [1].

Despite these many benefits, the use of AI in marketing is not without challenges. Issues such as data privacy, the need for large datasets for training, and the computational demands of training complex models are significant concerns [18]. Moreover, the interpretability of AI models remains a critical hurdle, particularly in sectors where understanding the decision-making process is crucial [19].

2.2 Facial Emotion Recognition (FER) in Consumer Behavior Analysis

FER-enabled systems offer an innovative approach to bridge the gap between emotional responses and data-driven decision-making. By analyzing non-verbal cues, these systems provide valuable information for applications ranging from personalized marketing to adaptive human-computer interactions. For example, Gaffary et al. [20] demonstrated how kinesthetic and facial expression displays can enhance emotion recognition, particularly for emotions with high activation levels through visuohaptic feedback. Similarly, Huang and Romano [21] explored artistic installations integrating shape-changing textiles and heart rate feedback to regulate emotions mindfully.

Advanced computational models further contribute to real-time emotion analysis. For example, Cohen et al. [22] advanced automatic FER by testing Bayesian network classifiers and introducing hidden Markov models for video-based facial expression segmentation. Sandiwarno et al. [23] introduced SES-Net, a multi-task deep neural model that simultaneously learns emotion and semantic information in e-learning environments. These studies collectively illustrate the transformative potential of emotion recognition technologies in enhancing user experiences across various domains.

FER could help analyze non-verbal consumer cues and can provide useful information about emotional responses that are elicited by advertisements and marketing [24]. The ability to decode emotions from facial expressions can enable marketers to evaluate the effectiveness of their campaigns and make the necessary adjustments to resonate more effectively with target audiences [16].

At the core of FER systems are advanced convolutional neural network architectures, such as Xception and ResNet-50 [4], which have previously proven highly effective in classifying complex emotional states. These architectures leverage large-scale datasets to train deep learning models that are capable of identifying consumer interest levels with impressive accuracy.

Despite these advancements, FER models face several challenges that limit their applicability across diverse demographics and cultural contexts. A major issue comes from the limited diversity in training datasets, which often fail to adequately represent variations in age, gender, ethnicity, background, and cultural expressions [25]. This lack of representation can lead to biased predictions, which reduces the reliability and fairness of FER models in real-world applications [26]. Addressing these biases requires the development of inclusive datasets that capture the richness of human expressions across different populations [27].

Ethical concerns surrounding FER also present significant barriers to adoption [18]. The collection and analysis of facial data raise privacy issues, especially when consumers are unaware or have not provided explicit consent for their data to be used. Furthermore, there is potential for FER technologies to be misused such as for manipulative purposes [28]. Transparency and user consent must be prioritized to ensure that FER applications align with ethical standards [29].

Another challenge lies in the interpretability of FER models [30]. While CNNs excel at identifying patterns in facial expressions, their “black-box" nature often makes it difficult to provide clear explanations of how decisions are made. This lack of transparency can hinder trust and adoption, particularly in contexts where interpretability is crucial.

Advancements in FER are further hindered by technical constraints, such as the need for high-quality input data and robust pre-processing pipelines [31]. Variations in lighting, camera angles, and background noise can impact the accuracy of FER models; so to overcome these challenges,

researchers are looking for novel techniques such as transfer learning and domain adaptation to improve model robustness and reliability in dynamic environments [32].

2.2.1 Facial Emotion Recognition (FER) in Marketing Research

Cohen et al. [22] presented significant advancements in automatic facial expression recognition from continuous video input. Their study tested various Bayesian network classifiers and introduced hidden Markov models (HMMs) for segmenting and recognizing facial expressions from video sequences. Results of their study indicated that Tree-Augmented Naive Bayes (TAN) classifiers outperformed Naive Bayes classifiers by effectively modeling dependencies among facial motion features.

In another study, Derbaix [33] explored how affective reactions to television advertisements influence attitudes towards the advertisement (Aad) and post-exposure brand attitude (Abp). They used facial expressions and traditional verbal measures to gauge affective reactions in a study involving 228 participants. Findings showed that verbal affective reactions significantly contributed to Aad and Abp.

In the research by Woltman Elpers et al. [34], the authors examined how moment-to-moment (MTM) entertainment value (EV) and information value (IV) influence consumers' likelihood of continuing to watch television commercials. Through two experiments, they found that high MTM EV decreased the likelihood of viewers stopping, while high MTM IV increased it. The interaction of high EV and IV further escalated the likelihood of viewers discontinuing viewership. These findings emphasize the importance of balancing entertainment and information in television commercials to avoid consumer overload and maintain viewer engagement.

Zhou et al. [35] addressed the challenge of balancing facial privacy with the need for facial information in business contexts. The authors propose the contour-as-face (CaF) framework, transforming face images into contour images that incorporate both non-outline and outline features of facial parts. Through three empirical studies, they compare human perceptions of face and contour images across 15 marketing-relevant dimensions and investigate the framework's effectiveness in

protecting anonymity related to identity, age, and gender. Thus, the results show that the CaF framework preserves critical perceptual information while making it nearly impossible to infer identity and very difficult to infer age and gender, thus resolving the privacy-perception trade-off.

Vakratsas and Ambler [36] conducted a comprehensive review analyzing over 250 articles and books to understand how advertising influences consumers. They challenged the traditional “hierarchy of effects" model and proposed a new framework categorizing advertising effects into intermediate (beliefs and attitudes) and behavioral effects (purchasing behavior). Their review identified various model types and provided five generalizations, including the significance of cognitive, affective, and experiential effects in advertising.

2.2.2 Challenges in Facial Emotion Recognition

Facial Emotion Recognition (FER) has garnered extensive research interest due to its applications in human-computer interaction, psychological analysis, and marketing [25]. Traditional FER systems rely on large datasets of facial expressions to train models capable of recognizing a range of emotions [37]. However, these datasets often lack diversity in demographics, cultural contexts, and emotional expressions, leading to models that perform inadequately when exposed to new content or different populations [38].

Datasets like CK+ [39] and JAFFE [40] are widely used but have limitations in terms of demographic diversity and expression variability [41]. Overfitting to specific datasets can result in biased models that do not capture the variability of facial expressions across different cultures, ages, and contexts [42]. This limitation emphasizes the need for more diverse and representative datasets to improve the generalizability of FER models [43].

2.3 Data Augmentation and Generalizability in Vision Deep Learning

Data augmentation has long been a foundational and most useful technique in machine learning, to improve model robustness and generalizability by artificially enhancing dataset diversity

and amount [44]. For FER, traditional data augmentation methods such as geometric transformations, such as rotations, flips, and scaling, and photometric adjustments, such as brightness, contrast, and saturation changes, have been widely used to mimic real-world variations in facial expressions and environmental conditions [45]. These techniques help FER models become more resilient to slight variations in input data, thus reducing overfitting and improving performance on unseen samples [46].

However, traditional augmentation methods often fall short in capturing the real-world variability [47]. For example, geometric transformations cannot generate entirely new facial expressions, nor can photometric adjustments account for variations in skin tone and lighting condition or cultural nuances in emotional expression [48]. This limitation poses challenges for FER systems, especially when applied to diverse situations where soft emotional cues and demographic differences play a crucial role [49].

To address these limitations, advanced generative techniques such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have transpired as powerful tools for augmenting FER datasets [50]. Unlike traditional augmentation methods, GANs and VAEs can generate completely new data samples that show complex and realistic variations in facial features and expressions [51].

For instance, GANs, especially architectures like StyleGAN2 from NVIDIA, excel at synthesizing high-resolution facial images with controllable attributes like age, gender, ethnicity, and emotional expression [52, 53]. These synthetic images can fill gaps in underrepresented categories and possibly increase the diversity and balance of training datasets [54].

VAEs, on the other hand, offer a probabilistic approach to data generation and learning latent representations of input data to create variations that align very closely with the original distribution [55]. This capability makes VAEs especially useful for generating very small and natural-looking variations in facial expressions that capture the nuances that are critical for precise FER [56].

By augmenting real-world datasets with synthetic samples, FER models can learn to recognize emotional expressions across a broader range of situations and demographics [27]. For example, based on previous research, combining synthetic data with real-world datasets has been shown to reduce overfitting, improve accuracy on previously unseen data, and increase the ability of FERs to generalize across diverse populations [57].

Despite these advancements, the use of generative techniques for data augmentation is not without challenges. One important concern is ensuring the quality and authenticity of synthetic data because poorly generated samples can introduce noise and bias into the training process and possibly undermine the performance of FER models [58]. Additionally, GANs and VAEs require substantial computational resources for training, which makes their deployment costly and time-intensive for many organizations [59].

Another important challenge is the potential for synthetic data to amplify existing biases in training datasets [60]. If the original dataset used to train a GAN or VAE is biased, the generated data will likely inherit and propagate those biases, perpetuating disparities in model predictions [61]. For instance, if the training data lacks representation from certain ethnic groups or age ranges, the synthetic data may fail to accurately reflect those populations, and thereby limit the inclusivity and fairness of FER systems [62].

To mitigate these issues, researchers have explored techniques such as fairness-aware data generation and adversarial debiasing to ensure that synthetic data contributes to more equitable and reliable FER models [63]. Additionally, validation processes, including cross-validation with real-world datasets and performance benchmarking, are essential to assess the impact of synthetic data on model outcomes [64].

2.4 Generative AI for Dataset Expansion in Vision Deep Learning

Generative AI has emerged as a transformative solution to the data challenges faced by vision deep learning, more particularly FER systems, by addressing critical limitations such as data

scarcity, imbalanced datasets, and demographic underrepresentation [27]. By generating synthetic data that supplements real-world datasets, generative models, particularly generative adversarial networks (GANs), can enhance the robustness and generalizability of FER systems [27]. This approach has proven valuable in mitigating overfitting and improving model performance across diverse applications [65].

GANs have demonstrated remarkable efficacy in producing high-quality, photorealistic facial images that capture subtle variations in expressions, lighting, and demographic features [66]. These synthetic datasets serve as an essential resource for training FER models to enrich their capacity to recognize a broader range of emotional expressions. For instance, StyleGAN2 can generate synthetic faces with diverse attributes, such as age, ethnicity, and gender, which are often underrepresented in real-world datasets [67]. This diversity is critical for improving the fairness and inclusivity of FER systems, particularly in applications where accurate emotion recognition across demographic groups is essential [30].

Combining synthetic data with real-world datasets, such as annotated social media reaction videos, further can enhance the generalizability of FER models [27]. Real-world data captures the complexity and variability of human expressions in dynamic environments, while synthetic data fills gaps in underrepresented categories and this hybrid approach allows FER models to achieve higher accuracy and robustness by leveraging the strengths of both data sources [27].

Despite these advancements, the use of synthetic data in FER raises several ethical and practical concerns [68]. One of the primary issues is data authenticity and synthetic data, while visually convincing, lacks the grounding in real-world experiences that real data provides and it may lead to biases in FER models, especially if the synthetic data does not accurately reflect the diversity and complexity of genuine human emotions [69]. For example, synthetic datasets may over-represent idealized or exaggerated facial expressions, and this results in skewing model predictions in real-world applications.

Biases in synthetic data generation also pose a significant challenge. GANs and other generative models are only as unbiased as the datasets they are trained on, and if the training data lacks

diversity or contains inherent biases, these biases are likely to be replicated in the synthetic outputs and it can perpetuate inaccuracies in FER models when deployed in sensitive scenarios such as hiring processes or security systems.

The ethical implications of using synthetic data extend beyond bias to concerns about transparency and accountability. Users of FER systems may not be aware that synthetic data has been used in model training, raising questions about informed consent and the ethical use of data [28]. Moreover, the potential misuse of synthetic data, such as in creating deepfakes or manipulative media emphasizes the need for robust regulatory frameworks to govern its application [70].

To address and solve these problems we need to adopt responsible AI practices [71]. Researchers and practitioners must clearly document the sources, methods, and limitations of synthetic datasets to ensure accountability. Additionally, incorporating fairness audits and bias detection mechanisms into the generative process can help mitigate potential issues and improve the ethical alignment of FER systems [72].

2.5 Virtual Reality (VR) and E-Commerce Business

Virtual reality (VR) has the potential for significant impacts on e-commerce; as VR offers immersive environments that revolutionize the way consumers engage with products and brands [73]. Unlike traditional shopping experiences, which are often limited by physical or digital constraints, three-dimensional VR provides a platform where consumers can interact with products in lifelike simulations [74]. VR allows consumers to explore products more intuitively and realistically [75].

The integration of VR into business strategies offers unprecedented opportunities for brands to differentiate themselves in highly competitive markets [76]. Researchers have explored various applications of VR in marketing [77], including virtual showrooms [78], interactive product trials [79], and immersive storytelling [80]. These studies highlight the technology's potential to enhance consumer perceptions and drive purchase intent [81].

The emotional engagement facilitated by VR is another significant advantage [82]. Immersive experiences create a sense of presence [81], where users feel as though they are physically situated

within the virtual environment [83]. This heightened sense of presence has been linked to increased satisfaction and stronger emotional connections with brands [84]. For example, virtual tours of real estate properties or 360-degree views of luxury items can evoke positive emotional responses and enhance brand perception and trust [85].

VR also enables personalization in e-commerce by tailoring experiences to individual preferences and needs [86]. Through AI-driven algorithms, VR platforms may be able to adapt product recommendations based on user data to create a unique and memorable shopping journey [87]. This level of customization is particularly valuable in competitive markets where differentiated and engaging customer experiences are crucial for retaining loyalty [88].

Additionally, VR can provide valuable data for businesses. User interactions within VR environments generate rich behavioral information such as gaze patterns, time spent on specific products, navigation paths, and stop times [89]. Understanding which virtual features capture the most attention can inform real-world commerce and advertising decisions [90].

Despite its potential, the integration of VR in e-commerce has many challenges. For example, the high cost of VR hardware and the technical expertise required to develop and maintain VR platforms remain significant barriers, particularly for small- and medium-sized enterprises [91]. Furthermore, ensuring a precise and intuitive user experience is critical, as poorly designed VR environments can eventually lead to frustration and disengagement [92]. Factors such as motion sickness and system latency must be carefully addressed to maximize user satisfaction and adoption rates [93].

Ethical considerations also play a critical role in the adoption of VR in e-commerce because businesses collect vast amounts of user data through VR interactions and ensuring data privacy and security is important [94]. Transparent policies and robust cybersecurity measures are essential to building consumer trust and avoiding potential misuse of sensitive information [95].

2.6 Challenges in AI-Powered E-Commerce Research

There are several significant challenges in the adaptation of e-commerce with AI. These challenges span technical, ethical, security, and operational dimensions, and emphasize the complexity of deploying AI technologies like facial emotion recognition and virtual reality in business strategies.

FER models often struggle with generalizability across diverse demographics, leading to biased predictions that can undermine their reliability and fairness [96]. This issue arises from the lack of diversity in training datasets, which frequently overrepresent certain demographics while underrepresenting others, such as minority ethnic groups, older adults, or individuals with typical facial expressions [27]. Consequently, these biases propagate into FER systems and result in skewed emotional analyses that fail to reflect real-world variability.

Also, the deployment of FER and generative AI in e-commerce raises critical ethical concerns, particularly related to data privacy and user consent [97]. Facial data is highly sensitive and must be collected, stored, and analyzed in ways that prioritize transparency and user rights [98]. However, many existing practices lack sufficient safeguards, exposing consumers to potential misuse of their biometric data [99]. Furthermore, synthetic data generation through GANs introduces additional ethical dilemmas, such as the risk of creating deepfake content or reinforcing biases embedded in the original datasets [100, 101]. Establishing robust ethical frameworks and regulatory guidelines is essential to mitigate these risks and build trust in AI-driven e-commerce systems [102].

Training deep learning models for real-time FER and VR applications requires substantial computational resources, including powerful GPUs, high memory capacity, and extensive databases and datasets, among many other requirements [31]. These requirements present a significant barrier for smaller organizations that may lack the infrastructure or budget to invest in such technologies. Additionally, the real-time nature of e-commerce interactions necessitates low latency and high accuracy predictions which further compounds the computational demands [103].

While FER and VR individually offer immense potential for e-commerce, their synergistic application remains underexplored. The integration of FER with VR environments could enable

marketers to analyze emotional responses to immersive shopping experiences in real-time and provide valuable information about consumer behavior. However, challenges such as aligning FER outputs with VR interactions, managing data synchronization, and ensuring precise user experiences must be addressed to unlock this potential. Current research has yet to comprehensively explore these areas, leaving a critical gap in understanding how FER and VR can be effectively combined to enhance e-commerce strategies.

2.7 Gaps in Literature

Despite substantial advancements in AI-driven e-commerce research and its application in VR environments, several critical gaps persist that highlight opportunities for further exploration and innovation.

Existing FER models lack the robustness required for diverse real-world applications, particularly in multi-cultural and multi-context environments [104]. Current research often focuses on controlled laboratory settings, which fail to capture the variability and unpredictability of real-world scenarios [105]. This gap emphasizes the need for more adaptable FER models that can generalize effectively across demographic and geographic differences.

While generative AI technologies such as GANs have proven effective in enhancing FER capabilities, their ethical implications require deeper investigation [28]. Issues such as data authenticity and transparency in synthetic data usage [106] and the potential for misuse (e.g., deepfakes) remain underexplored [107]. Developing ethical guidelines and validation frameworks for generative AI applications in e-commerce is essential to ensure responsible deployment [29].

Also, limited research has been conducted to isolate the specific VR design elements and exposure sequences that impact long-term consumer engagement and purchase intent. While studies acknowledge the importance of interactivity and immersion [108], there is little consensus on how these factors interact with consumer preferences and cognitive load.

The interplay between FER insights and VR environments remains poorly understood. While FER excels at decoding emotional responses in static or video-based advertisements, its applica-

tion in novel, immersive VR settings is underdeveloped. Key questions, such as how VR design influences FER outputs require further investigation.

Another critical gap lies in understanding the long-term effects of VR shopping experiences on consumer behavior [109]. While current research focuses on immediate engagement and purchase intent, there is a limited exploration of how VR interactions influence consumer satisfaction over time. Examining these longitudinal impacts could offer a more comprehensive view of VR's role in e-commerce.

The potential of integrating multiple data modalities, such as FER, eye-tracking, and physiological signals, into a unified framework for analyzing VR shopping behavior is largely untapped [110]. Multi-modal approaches could provide richer and more holistic insights into consumer preferences, enabling the design of more personalized and effective marketing strategies [111].

This dissertation seeks to address some of these gaps by integrating FER using CNNs and improving its generalizability by generative AI data augmentation and interacting VR technologies into a cohesive framework for AI-driven e-commerce strategies. By developing more generalizable FER models, exploring ethical approaches to synthetic data generation and privacy concerns related to data collection, and advancing the integration of FER and VR, this research aims to contribute to the development of robust, inclusive, and ethically responsible business marketing methodologies that transform the e-commerce landscape.

Chapter 3

Data Overview

This dissertation leverages a variety of data sources and experimental designs to train, validate, and test the vision deep learning models, as well as to investigate consumer engagement and purchase intent in virtual advertising environments. Each research aim utilizes a distinct dataset and methodology designed for its specific objectives. A detailed description of the data and analytical methods used for each research question is provided in the respective results chapters. However, an overview of the data used across this dissertation is summarized in Table 3.1.

Table 3.1: Data Used for Each Research Aim

Research Aim	Data Source	Sample Size
Aim 1	NeuroBioSense Dataset	10,450 images
Aim 2	NeuroBioSense Dataset Synthetic Image(StyleGAN2) Dataset YouTube Reaction Dataset	10,450 images 2,022 images 2,000 images
Aim 3	Desktop VR Study Surveys Dataset	55 participants
Aim 4	Desktop VR Study Reaction Videos Dataset	443,383 images

In summary, Research Aim 1 (effectiveness of CNNs in predicting consumer engagement) and Research Aim 2 (addressing dataset limitations with generative AI and social media data) both used the NeuroBioSense data [112], which is a dataset publicly available on Mendeley and is comprised of video data of facial expressions for 58 participants while they watch various advertisements. Additionally, Research Aim 2 integrated two additional datasets that were collected as part of this dissertation, a synthetic image dataset that was generated using StyleGAN2 and video frames collected and exported from YouTube.

For Research Aim 3 (effects of VR design complexity and exposure sequencing on engagement) and Research Aim 4 (extending vision deep learning models to VR), a desktop VR study

was performed with 55 participants from San Jose State University (SJSU) and data collected from that study were used to answer these research aims. Research Aim 3 focused on the self-reported survey responses and Research Aim 4 evaluated the facial expression video data of the participants. The study had approval for human subjects research from the Colorado State University IRB (protocol #6032).

These experimental designs and datasets provide a robust empirical foundation for exploring the intersection of AI-driven marketing personalization and immersive technologies. Further details on the methodologies and findings are elaborated in the next chapters.

Chapter 4

Aim 1: Effectiveness of CNNs in Predicting Consumer Engagement

4.1 Aim 1 Summary

Advances in artificial intelligence (AI), specifically convolutional neural networks (CNNs), have significantly enhanced the ability to analyze and interpret consumer behaviors in digital advertising. This chapter employs two CNN architectures, Xception and ResNet-50, to decode consumer interest by analyzing facial expressions captured during advertisement viewing. The primary goal is to assess how these models can predict consumer engagement and provide actionable insights that could refine marketing strategies and increase customer interaction. Utilizing a comprehensive dataset, which includes videos of participants' faces while watching various advertisements, the chapter demonstrates the effectiveness of these models in distinguishing between interested and not-interested. The analysis reveals that certain emotions, prominently happiness, are strong indicators of consumer interest, while emotions like disgust and fear correlate with disinterest. The research findings suggest that CNNs, particularly the Xception model, offer substantial advantages in recognizing these patterns, thereby presenting a transformative approach for marketers to understand and predict consumer behavior dynamically. This not only aids in personalizing advertising content but also enhances the overall efficacy of digital marketing campaigns.

4.2 Introduction

Digital transformations have notably shifted consumer shopping behaviors, leading to significant changes in e-commerce [113]. The convenience of online shopping has not only recalibrated consumer expectations [114] but has also mandated traditional retail outlets to innovate, such as by offering unique, visually captivating experiences to maintain customer loyalty [115].

Within this evolving commercial environment, the nuanced interplay between consumer emotions and decision-making processes emerges as a primary point of interest for research and application [116].

Recognizing and interpreting facial expressions helps understand consumer behaviors, given their significant impact on preferences and decisions [117]. Research highlighting the influence of non-verbal cues, such as those by Shiv and Fedorikhin [118] and Mehrabian et al. [119], demonstrate the necessity for advanced facial expression recognition technologies. These technologies are pivotal not only in enhancing customer interactions, but also in their application across various sectors, including interactive gaming, sociable robotics, and especially in data-driven marketing strategies [120].

Despite notable advancements, the challenge of accurately identifying and interpreting complex human emotions through AI remains substantial [121]. The development and refinement of AI-based systems, particularly those employing machine learning and deep learning techniques, are crucial in overcoming these barriers [7]. Deep learning, with its capacity for feature extraction and classification through multi-layered neural networks, has revolutionized the field of computer vision [122], offering substantial improvements in facial expression recognition tasks [123].

The emergence of convolutional neural networks (CNNs), introduced by LeCun et al. [124] and significantly advanced through architectures like AlexNet [125] and GoogLeNet [126], have been crucial in image classification. These technologies facilitate the categorization of images into predefined labels, optimizing the training process to improve accuracy and efficiency in interpreting facial expressions. Despite these technological strides, the application of CNNs in evaluating consumer engagement with advertisements remains under-explored, presenting a notable gap in current research.

This chapter aims to address this gap by leveraging deep learning, specifically convolutional neural networks, to advance the understanding of consumer engagement in the context of digital advertising. The objective of this chapter is to demonstrate the effectiveness of using CNNs to predict consumer behaviors. By examining the impact of various factors, such as advertisement content

on viewer engagement through facial expression analysis, this research seeks to provide actionable insights for the creation of more effective and personalized marketing strategies. Specifically, it addresses the need for innovative approaches in capturing and analyzing consumer reactions, contributing to the development of personalized and emotionally impactful advertising content in the digital marketplace.

4.2.1 Problem Statement

Previous research has shown the value of AI and facial recognition in marketing research. However, recent advancements in deep learning, particularly CNNs, and computational resources have created a gap in knowledge regarding the effectiveness of CNNs in predicting consumer behavior. In this chapter, we seek to address this gap by answering the following research questions:

1. **RQ 1.1:** How effective are two prominent convolutional neural network (CNN) architectures, Xception and ResNet-50, in distinguishing consumer engagement (interested vs. disinterested)?

2. **RQ 1.2:** What role do specific emotional cues (e.g., happiness, disgust, fear, anger, etc.) play in consumer interest classification?

3. **RQ 1.3:** What are the implications of facial expression analysis findings for personalized marketing strategies in digital advertising?

4.3 Analytical Methodology

This section describes the CNN framework used in this research. Convolutional neural networks (CNNs) are fundamental to advancing machine learning applications in image recognition and processing. This section discusses the architectures and functionalities of two prominent CNN models: ResNet-50 and Xception. Each model leverages unique mechanisms to optimize performance and accuracy in tasks such as facial feature extraction, essential for analyzing consumer engagement in advertising.

4.3.1 Convolutional Neural Network Architecture

There were two CNN models trained and compared for their ability to predict consumer interest based on video data of participants watching advertisements. Each CNN model used a different architecture, which was ResNet-50 and Xception.

ResNet-50

ResNet-50 (Residual Network with 50 layers) is a widely used CNN architecture in deep learning [127]. ResNet-50 is part of the broader family of ResNets that revolutionized the field by addressing the problem of vanishing gradients in very deep networks. The central innovation of ResNet-50 is the introduction of residual learning, which involves the use of shortcut connections, or skip connections, that bypass one or more layers. These connections allow the network to learn residual functions concerning the layer inputs. This approach mitigates the problem of vanishing gradients, enabling the training of much deeper networks.

ResNet-50 employs a bottleneck architecture to enhance computational efficiency and reduce parameters, using three convolutional layers in each bottleneck block: 1x1 (reduces dimensionality), 3x3 (performs spatial processing), and 1x1 (restores dimensionality) convolutions. This design maintains high accuracy while being efficient. The architecture comprises 50 layers, including convolutional, batch normalization, and rectified linear unit (ReLU) activation layers, organized into five stages with increasing filter numbers, demonstrating exceptional performance. Thus, in this research, we use ResNet-50 as one of the CNN architectures due to its high accuracy and manageable computational demands. Figure 4.1 visualizes the ResNet-50 structure.

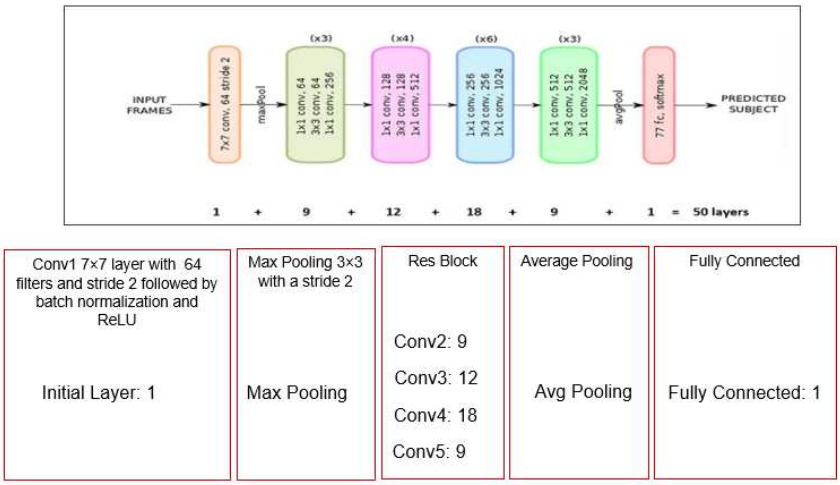


Figure 4.1: ResNet-50 layer structure.

Xception

Xception is a type of CNN architecture that was introduced by François Chollet [128], the creator of the Keras deep learning library. Xception stands for “extreme inception," and it builds upon the idea of inception modules introduced in the GoogLeNet architecture. However, Xception takes a different approach to the convolutional layers within these modules [128].

The key innovation of Xception is the use of depthwise separable convolutions, which decompose the standard convolution into two separate operations: depthwise convolution and pointwise convolution [129]. This separation significantly reduces the number of parameters and computational complexity compared to traditional convolutions while retaining expressive power [130].

By adopting depthwise separable convolutions, Xception aims to achieve better efficiency and performance on image classification tasks and it has been shown to outperform previous state-of-the-art architectures on various benchmarks while being computationally more efficient [131].

Xception employs a combination of architectural advancements alongside techniques like batch normalization and ReLU activation to enhance model stability and accelerate convergence, and these components contribute to Xception’s exceptional performance in image classification tasks [132].

Xception represents an important advancement in CNN architectures; it demonstrates the effectiveness of depthwise separable convolutions and it facilitates more efficient and powerful models in the field of computer vision [131]. Figure 4.2 visualizes the Xception structure.

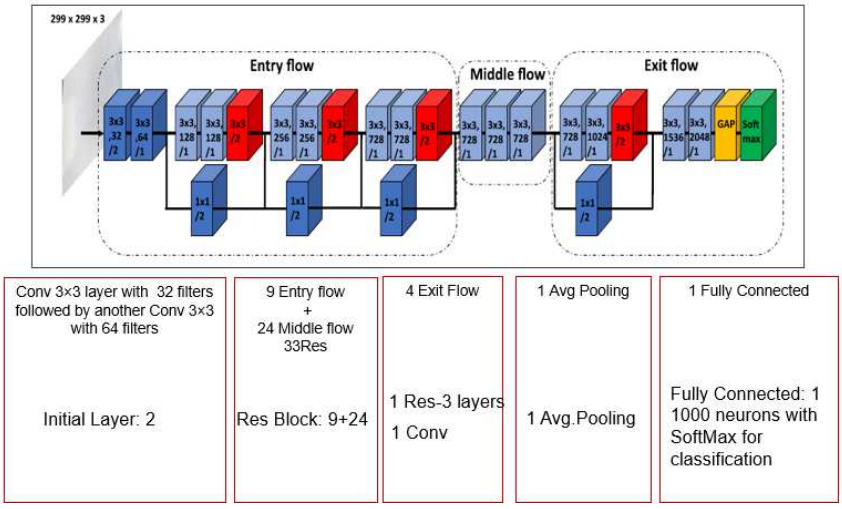


Figure 4.2: Xception layer structure.

4.3.2 Basic Convolutions Neural Network Components

CNN models, despite their diverse variations, share a core structure that follows a consistent pattern, which includes an input layer, alternating convolutional and pooling layers, followed by one or more fully connected layers, scattered with activation functions, and concludes with an output layer [133]. The initial part of the network functions as a feature extractor, using convolution and pooling layers in succession to process and transform raw input into abstract, higher-level features. The fully connected layers, combined with activation functions, then undertake tasks like classification based on these features. To enhance performance, CNNs also incorporate regulatory mechanisms such as batch normalization and dropout, alongside various mapping functions. Figure 4.3 represents an illustration of convolutional neural network architecture for facial feature extraction.

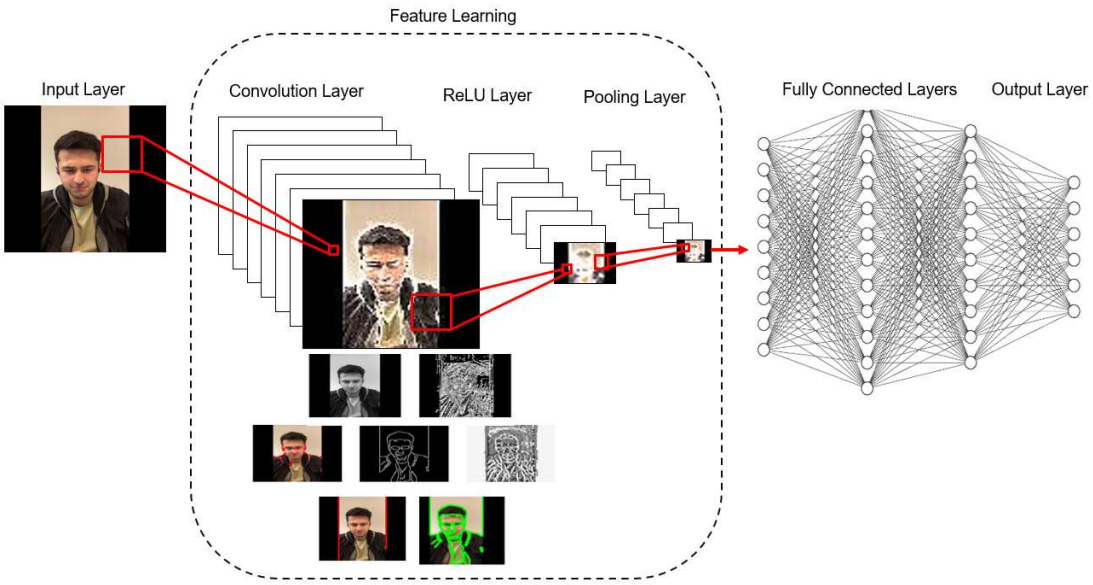


Figure 4.3: A proposed CNN architecture for engagement detection classification.

Convolutional Layer

Convolutional layers operate by applying multiple filters (kernels) to an image to extract features such as edges, textures, or patterns [134]. These filters are matrices of weights that slide (convolve) across the image in small steps, a process called stride. At each position, the filter performs an element-wise multiplication with the part of the image it covers, and the results are summed up to produce a single pixel in the output feature map. This operation is repeated across the entire image, creating a feature map for each filter.

The convolution operation in a convolutional layer can be mathematically represented as:

$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(m,n) \cdot K(i-m, j-n) \tag{4.1}$$

where $S(i,j)$ is the output of the convolution at position (i,j), I is the input image, K is the kernel or filter, and $*$ denotes the convolution operation. The sums over m and n iterate over the entire kernel, applying the kernel's weights to the input image's pixels to produce the output feature map.

Activation Layer (ReLU)

The activation layer, particularly employing the Rectified Linear Unit (ReLU), is a pivotal component in convolutional neural networks that introduces non-linearity into the model. ReLU is defined as $f(x) = max(0, x)$, meaning it outputs the input directly if it is positive, otherwise, it will output zero. This simple yet effective function allows CNNs to learn complex patterns and relationships in the data [135]. Positioned after convolutional layers, ReLU helps in mitigating the vanishing gradient problem, where gradients become too small for the network to learn effectively, by providing a consistent gradient for positive inputs. Thus, its computational simplicity and efficiency in accelerating the convergence of stochastic gradient descent make ReLU a preferred choice in deep learning architectures [136]. By enabling CNNs to model non-linear relationships without significant computational overhead, ReLU layers enhance the network's ability to perform tasks like image recognition and classification with high accuracy.

Dropout Layer

The dropout layer is a regularization technique used in convolutional neural networks to prevent overfitting and it helps the model generalize better to new data [137]. It works by randomly deactivating a fraction of the input units in each training step, which forces the network to learn more robust features that are not dependent on specific neurons [138]. This reduces the model's sensitivity to particular weights, enhancing its ability to generalize. Dropout is especially useful before fully connected layers, which are prone to overfitting due to their high parameter count and it is an important hyperparameter that helps manage the trade-off between underfitting and overfitting. Overall, the dropout layer is an effective tool for improving the robustness and performance of CNN models.

The dropout layer randomly deactivates a proportion of neurons in the network during training, described by:

$$y_i = x_i \cdot D_i \tag{4.2}$$

where x_i is the input to a neuron, y_i is the output, and D_i is a random variable such that:

$$D_i = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } (1-p) \end{cases} \tag{4.3}$$

Here, p is the dropout rate, indicating the probability that any given neuron is set to zero, and $(1-p)$ is the probability of a neuron remaining active. This technique effectively reduces overfitting by preventing complex co-adaptations on training data.

Pooling (Down Sampling) Layer

The pooling layer reduces the spatial dimensions of the feature maps from the convolutional layers, simplifying the information and making the detection of features less sensitive to location and scale [139].

Max pooling is a common technique in convolutional neural networks that simplifies the output by selecting the maximum value from each region of the feature map, typically set with k=1, focusing on the most prominent features. This results in a condensed representation z, which retains crucial information from each feature map while reducing computational load and enhancing the model's generalization abilities. By extracting and emphasizing the most significant features, max pooling effectively aids in tasks like image classification, where recognizing key features is vital for categorizing images.

Max pooling can be described as:

$$S(i,j) = \max_{a,b \in \text{Region}(i,j)} I(a,b) \tag{4.4}$$

Here, $S(i,j)$ represents the output of the pooling operation at position (i,j), $I(a,b)$ is the input feature map, and Region(i,j) defines the local region around position (i,j) over which the max operation is applied. The max pooling operation selects the maximum value from each region of the input feature map, effectively downsampling the input while preserving the most significant features.

Fully Connected (Dense) Layer

The fully connected (dense) layer is an essential component of convolutional neural networks, acting as a bridge between feature extraction and classification stages. Unlike convolutional layers, where neurons connect to only a local input region, neurons in fully connected layers connect to all activations from the previous layer. These layers are responsible for interpreting features extracted earlier in the network and making predictions by learning non-linear combinations of these features. The final layer often uses a softmax function to convert these combinations into probabilities for classification tasks. Essentially, fully connected layers integrate all extracted features to drive the network's decision-making process, crucial for both classification and regression outputs.

A fully connected (dense) layer in a neural network can be mathematically represented by the equation:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{4.5}$$

where $\mathbf{x}$ is the input vector to the fully connected layer, $\mathbf{W}$ represents the weight matrix associated with the layer, $\mathbf{b}$ is the bias vector, and $\mathbf{y}$ is the output vector of the layer. In this context, each element of the output vector $\mathbf{y}$ is obtained by computing the weighted sum of the inputs plus a bias term, followed by an activation function (not shown in this equation). The fully connected layer plays a crucial role in neural networks by integrating learned features into predictions or classifications.

Softmax Layer

This layer plays a crucial role in the architecture of CNNs and it is typically positioned at the end of the network and functions as an activation layer that transforms the outputs of the previous layers into a probability distribution over predicted output classes. The softmax layer computes the exponential of each input value and normalizes these values by dividing by the sum of all exponentials; this ensures that the output values are in the range of 0 to 1 and sum to 1. This transformation provides a probabilistic interpretation of the network's outputs.

The formula for the softmax function, applied in the context of neural networks, especially as the activation function in the output layer for classification tasks, is defined mathematically as follows:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4.6}$$

where: z_i is the input to the softmax function for class i, e^{z_i} is the exponential of the input z_i, and the denominator $\sum_{j=1}^{K} e^{z_j}$ is the sum of the exponentials for all possible classes K, which normalizes the output.

4.4 Data Overview

This section presents an in-depth explanation of the data utilized in this chapter, predominately the NeuroBioSense dataset [112] used to train, validate, and test our CNN model for interested/not-interested. This section also describes the datasets employed for training the Facial Emotion Recognition (FER) model, which was used to identify underlying emotions in our interested/not-interested model. The NeuroBioSense dataset, central to this study, offers a diverse representation of consumer demographics and emotional responses to advertisements across various sectors. Additionally, the FER model leverages multiple well-established datasets to enhance its ability to accurately predict emotional states from facial expressions.

4.4.1 NeuroBioSense Dataset

The NeuroBioSense dataset [112] was used for this research. The data is based on a study of 58 participants ranging in age from 18 to 70, providing a diverse representation across consumer demographics. Participants were segmented into three groups, each exposed to advertisements from specific sectors: *(1)* cars and technology, *(2)* food and market, and *(3)* cosmetic and fashion. This segmentation aimed to uncover sector-specific emotional responses and consumer behaviors, providing insights into the varying impacts of advertising across these domains. Table 4.1 provides an overview of the NeuroBioSense study design.

Table 4.1: Demographics and Advertisement Categories in the NeuroBioSense Dataset

Number of Participants	58
Participant Age Range	18-70
Number of Unique Advertisements	35
Categories of Advertisements	
Car and Technology	20 Participants, 10 Ads
Food and Market	20 Participants, 10 Ads
Cosmetic and Fashion	18 Participants, 15 Ads

There were a total of 1,045 videos of participant faces while watching advertisements. A summary of the number of videos tagged as interested and not-interested by ad category is provided in Table 4.2.

Table 4.2: Summary of Interested and Not-Interested Videos by Ad Category

Ad Category	**Total Videos**	**Interested**	**Not Interested**
Car and Technology	302	251 (83.1%)	51 (16.9%)
Food and Market	337	186 (55.2%)	151 (44.8%)
Cosmetic and Fashion	406	265 (65.3%)	141 (34.7%)

A visual of the breakdown of ages by ad category and interested/ not-interested is shown in Figure 4.4. As evidenced by this boxplot, there was more spread in ages for participants that watched the food and market category, compared to the other two categories.

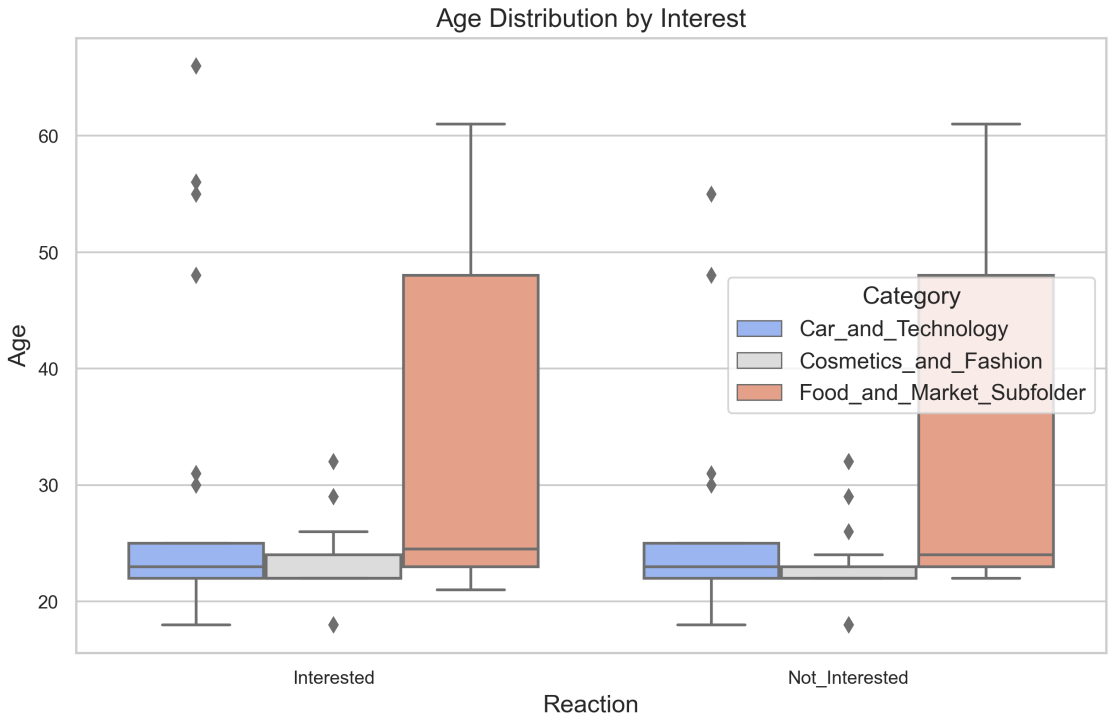


Figure 4.4: Distribution of participant ages based on interest levels across ad categories.

4.4.2 Dataset Used for Training the Facial Emotion Recognition Model

The Facial Emotion Recognition (FER) model, which leverages the Visual Geometry Group (VGG) architecture [140], has previously been trained on extensive datasets designed to capture a wide range of human emotions through facial expressions. A primary dataset used was FER-2013 [141], which consists of 35,887 grayscale images of faces, each measuring 48×48 pixels, and labeled with one of seven basic emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. In addition to FER-2013, other datasets like AffectNet, CK+ (Extended Cohn-Kanade Database), JAFFE (Japanese Female Facial Expression Database), and EmotiW (Emotion Recognition in the Wild) contribute to the model's robustness. AffectNet [142] provides around one million facial images from the Internet, labeled with both categorical and continuous emotion labels, enhancing the model's ability to generalize across diverse expressions and settings. CK+ [143] offers 593 video sequences from 123 subjects, showcasing transitions from neutral faces to peak expressions, while JAFFE [144] includes 213 images of Japanese female subjects, each displaying one of seven facial expressions. EmotiW [145], with its collection of images and videos from various sources,

presents a real-world scenario for emotion recognition. These datasets collectively ensure that the FER model is well-equipped to predict emotional states accurately across different populations and environments.

In our study, the FER model was used to identify the emotional states (i.e., angry, disgust, fear, happy, sad, surprise, neutral) of the participants from the videos of their faces while watching the advertisements. These emotional states were then correlated to the self-reported labeling of "interested" / "not interested".

4.4.3 Data Pre-Processing

Each video file was labeled as "interested" or "not interested" based on participants' self-reported responses after watching the advertisement. Additionally, each video was matched with metadata regarding the participants' age, gender, and advertisement category. The videos were not all the same duration, so 10 random frames were extracted from each video to ensure consistency across the dataset. Additionally, videos were collected using participants' smartphones, hence not all of the videos had the same dimensions. Padding was applied to some of the images to standardize the frames to a resolution of 300×300 pixels. The processed frames were then categorized into "interested" and "not interested" groups, facilitating the subsequent analysis and model training stages. Figure 4.5 illustrates this pre-processing pipeline.

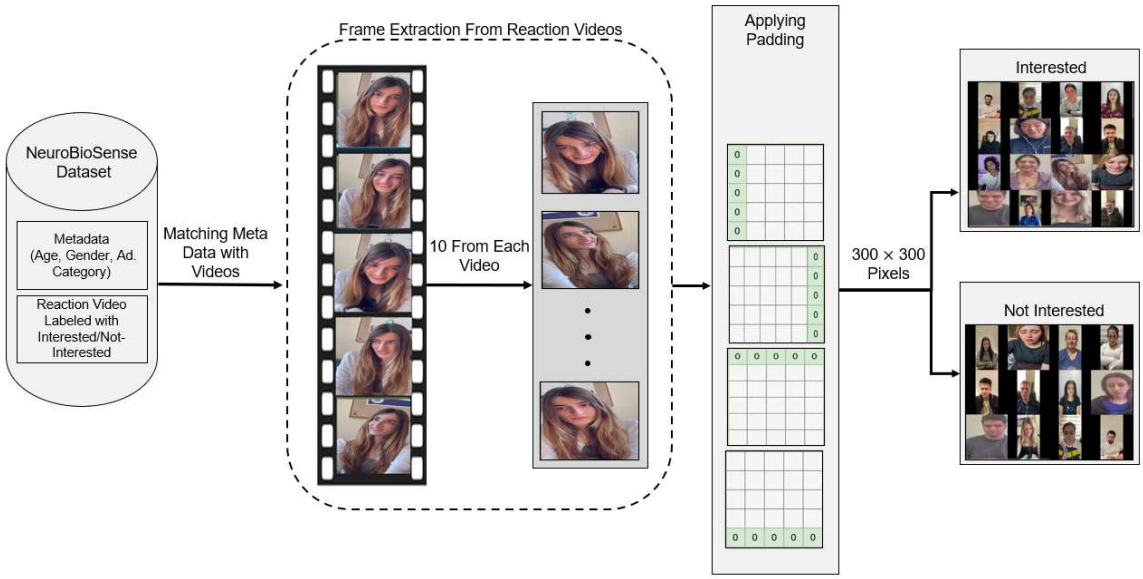


Figure 4.5: Data pre-processing pipeline used for analyzing video reactions.

There were a total of 10,450 frames (7020 "interested" and 3430 "not interested") used in the analysis. These frames were randomly allocated to different subsets for model training (3657 images), validation (1568 images), and testing (5225 images), see Figure 4.6. This approach ensured balanced representation across the different phases of model development.

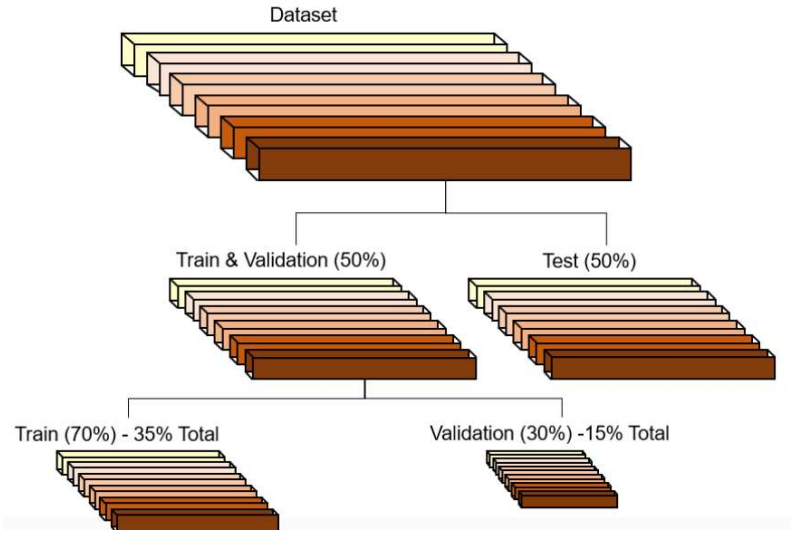


Figure 4.6: Distribution of data split for training, validation, and testing.

4.4.4 Fine Tuning and Data Augmentation

Adaptive Moment Estimation (Adam) is a widely adopted optimization algorithm in the area of computer vision and is crucial in the fine-tuning phase of training [146]. This algorithm incorporates principles from two established optimization techniques, namely adaptive gradient descent (AdaGrad) [147], which allocates distinct learning rates to each parameter of the model, and root mean square propagation (RMSProp) [148], which similarly assigns varying learning rates based on the average of prior magnitudes.

Data augmentation is an essential strategy in machine learning, enhancing the generalizability of predictive models, especially within the domain of image classification [149].

In this study, a dynamic data augmentation pipeline was implemented using Keras's ImageDataGenerator [150]. Image data was augmented on the fly during training, providing the model with a diverse array of inputs at each iteration (epoch). More particularly, each time the data is loaded from memory, a minor transformation is applied to the images, producing slightly varied data. Consequently, the model does not receive identical data in every epoch, reducing its susceptibility to overfitting. This technique is particularly advantageous when dealing with smaller datasets.

The augmentation protocol included random rotations of up to 30 degrees, simulating various camera angles and orientations of subjects. Furthermore, images were subjected to random horizontal flips, effectively doubling the directional variance of the dataset. A zoom range of up to 20% was utilized to mimic variations in image focus resulting from the camera's zoom function. Additionally, a shear transformation intensity of 0.2 was employed to introduce geometric distortions related to shifts in perspective.

The training process was conducted using Google Cloud's e2-highmem-16 instance, which includes a processor configuration of 16 vCPUs (8 cores) and 128 GB of RAM. For GPU acceleration, an NVIDIA Tesla P100 was utilized, ensuring efficient handling of the computational demands associated with model training and data augmentation.

4.5 Results and Discussion

4.5.1 Convergence of Binary Classifier Over Training Epochs

For validation purposes, understanding the convergence behavior of a binary classifier over training epochs is crucial in assessing the learning influence and stability of the model [151]. Convergence is an indication that the model is learning to generalize from the training data and is making consistent progress towards minimizing the error on unseen data [152]. In this research, the convergence patterns of our binary classifiers were closely monitored through the analysis of loss and accuracy metrics over 300 epochs. This analysis was performed on the training and validation data splits.

Model Loss

Loss is assessed using the loss function, which is a quantitative measure that captures the discrepancy between the predicted outputs of the model and the actual outcomes [153]. Since interested/ not-interested is a binary outcome, we used the Binary Cross Entropy loss function (Equation 4.7), where L is the loss function, N is the number of samples, y_i is the actual label of the ith sample and $\hat{y_i}$ is the predicted probability that the ith sample belongs to the positive class.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \times \log(\hat{y_i}) + (1 - y_i) \times \log(1 - \hat{y_i})] \tag{4.7}$$

The loss trajectory over the 300 training epochs is shown for Xception (left) and ResNet-50 (right) in Figure 4.7. In both plots, the declining loss curve over successive training epochs signifies that the model is effectively minimizing the prediction error. Our observations show a steady decrement in both training (blue line) and validation (orange line) loss, explaining the model's improvement in performance as learning progresses. The training loss provides insight into how well the model fits the training data, while the validation loss offers an indication of how well the model generalizes to new data. Notable is the Xception CNN's ability to minimize loss, indicating improved accuracy and successful generalization over time; whereas the ResNet-50 model has more initial fluctuations in the loss curve.

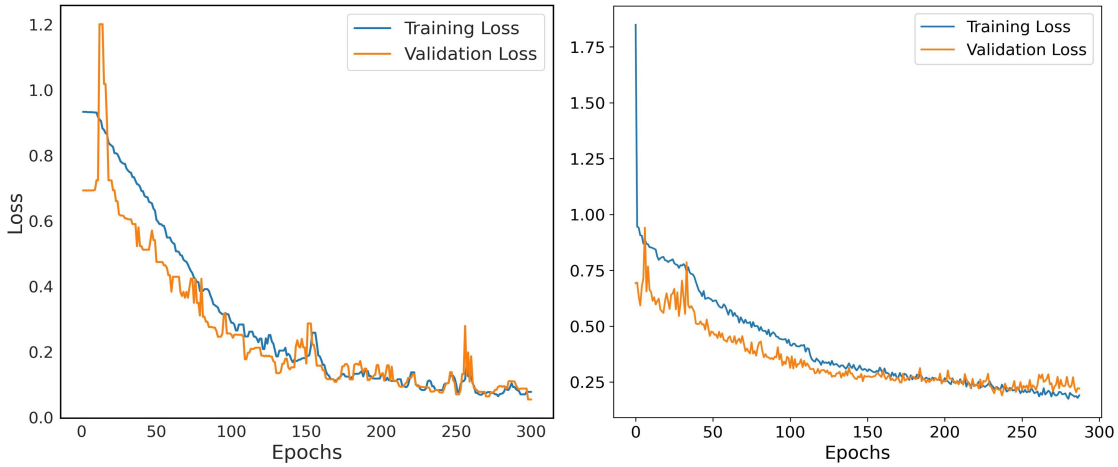


Figure 4.7: Loss curve for training (blue) and validation (orange) data for Xception (left) and ResNet-50 (right) models.

Model Accuracy

The model's accuracy reflects its proportion of correct predictions, as defined in Equation 4.8. Where N is the total number of predictions, $\hat{y}_i$ is the predicted label, y_i is the true label, and I is the indicator function that is 1 when the predicted label equals the true label and 0 otherwise.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i = y_i) \tag{4.8}$$

Figure 4.8 shows the training and validation accuracy for the Xception and ResNet-50 models over 300 epochs. The plots demonstrate the model's capacity to correctly classify the training data while also validating its proficiency on unseen data, indicating effective learning and generalization capabilities.

For the Xception accuracy, the upward trend in both training and validation accuracy indicates the model's enhanced predictive ability. Notably, the consistency between training and validation accuracy suggests that our model is not overfitting the training data but is instead learning generalizable patterns.

For the ResNet-50 accuracy, initially, both training and validation accuracy exhibit rapid improvement, reaching around 70% accuracy within the first 50 epochs. As training continues, the accuracy steadily increases, demonstrating the model's ability to learn and generalize from the

training data. In particular, the validation accuracy shows more fluctuations compared to the training accuracy, which is expected as the model is evaluated on unseen data. However, these fluctuations diminish as training progresses, indicating that the model is stabilizing and learning effectively. By the end of the training period, both training and validation accuracy converge around 90%, suggesting that the ResNet-50 model is performing well with minimal overfitting. This convergence indicates a good balance between fitting the training data and maintaining generalization capabilities which reflects a successful training process and the effectiveness of ResNet-50 for the given classification task.

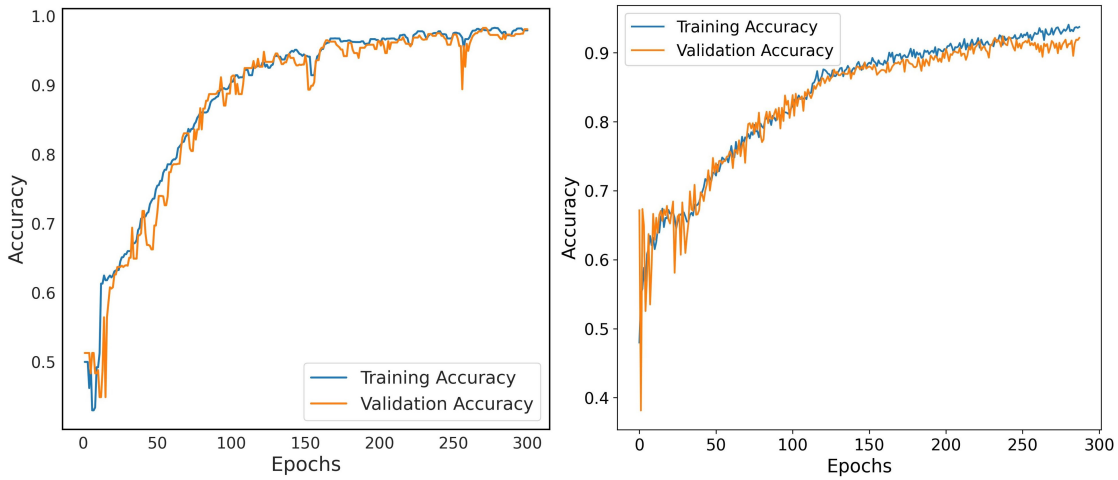


Figure 4.8: Accuracy curve for training (blue) and validation (orange) data for Xception (left) and ResNet-50 (right) models.

Overall, the convergence trends observed in our models are confirmed by the convergence theory in machine learning, which proposes that a well-tuned model, with sufficient data and under appropriate learning conditions, should exhibit a reduction in loss and an improvement in accuracy over time [154]. These trends confirm that our models, equipped with data augmentation and regularized through early stopping, are capable of learning and generalizing effectively, embodying the desired characteristics of a robust binary classifier.

4.5.2 Evaluation Metrics

Evaluation metrics critical for this study include precision and recall, Receiver Operating Characteristic (ROC) Curve, Confusion Matrix, and F1 score, reflecting the nuanced categorization of engagement levels. These analyses were conducted on the test split of the dataset.

Precision and Recall

Precision measures the accuracy of the positive predictions made by the classification model (Equation 4.9). It is the proportion of true positive results in all positive predictions made [155]. High precision indicates that the model is accurate in its positive predictions, but it does not indicate how many actual positives the model fails to detect. In this analysis, a true positive represents "interested" and a true negative represents "not interested."

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{4.9}$$

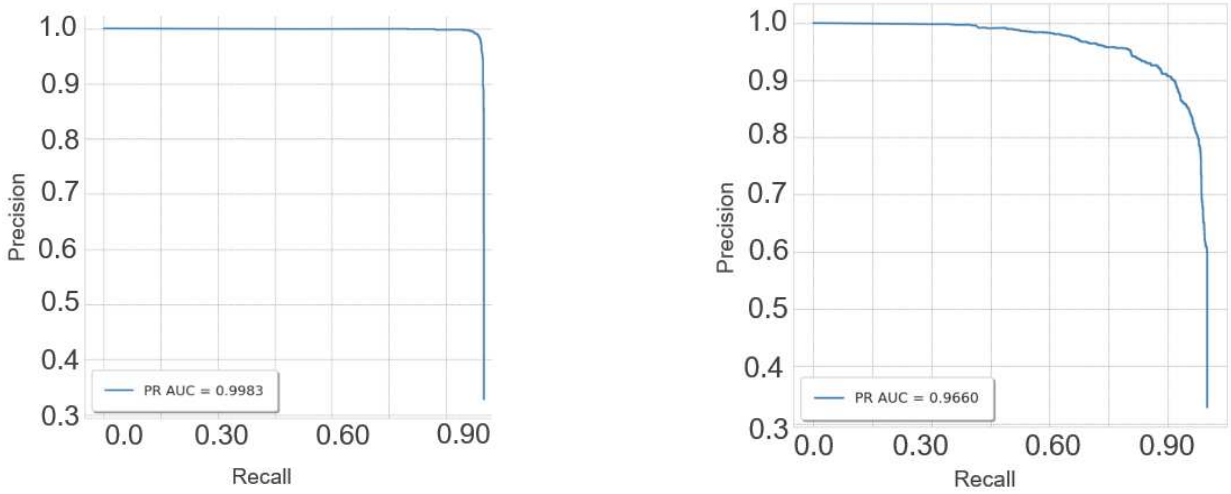
Recall (Equation 4.10) is sensitivity or true positive rate, and it measures the ability of the classification model to find all the relevant cases (positives). High recalls mean that the model is good at detecting the positive cases but it does not indicate how many negative cases are mistakenly identified as positive.

$$\text{Recall (TPR)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{4.10}$$

As such, the precision-recall (PR) curves show the trade-off between precision (proportion of positive identifications that were actually correct) and recall (proportion of actual positives that were identified correctly) for different threshold settings.

For the Xception model (Figure 4.9a), the PR curve shows an area under the curve (AUC) of 0.9983, which is very close to 1. This indicates excellent performance, as the model can maintain a high precision while also achieving a high recall. Similarly, for the ResNet-50 model (Figure 4.9b), the PR curve shows an AUC of 0.9660. Although slightly lower than the Xception model, this AUC still indicates strong performance, with the model maintaining a good balance between

precision and recall. Overall, both models demonstrate effective classification capabilities, with the Xception model slightly outperforming the ResNet-50 model in terms of precision and recall.



(a) Xception model performance

(b) ResNet-50 model performance

Figure 4.9: Precision-recall curves for the different models.

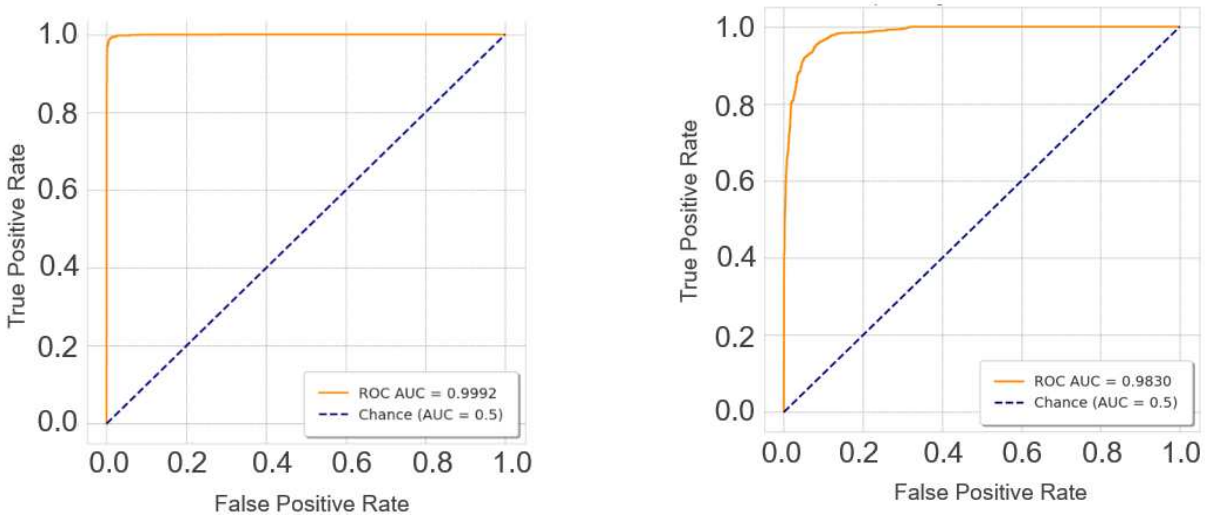
Receiver Operating Characteristic Curve

The ROC curve is another common tool for evaluating the performance of a binary classification system. It plots the true positive rate (TPR, or recall - Equation 4.10) against the false positive rate (FPR) at various threshold settings. Where FPR is defined as the number of incorrect positive predictions (False Positives) divided by the total number of negatives (true negative plus false positives), Equation 4.11.

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}} \tag{4.11}$$

As shown in Figure 4.10a, the ROC curve for our Xception model is very close to the upper-left corner, which denotes an excellent true positive rate for almost all thresholds. More specifically, the $AUC_{Xception} = 0.9992$. Similarly, as shown in Figure 4.10b, the ROC curve for the ResNet-50 model demonstrates a strong performance, with the curve nearing the upper-left corner. Where

$AUC_{ResNet-50} = 0.9830$ reflects its excellent capability to distinguish between the positive and negative classes.



(a) Xception Model ROC curve

(b) ResNet-50 Model ROC curve

Figure 4.10: Receiver Operating Characteristic (ROC) curves for the different models.

Confusion Matrix

A confusion matrix is another foundational tool in machine learning for visualizing the performance of classification models; it helps in understanding not just the overall accuracy of the model, but also provides detailed insights into the types of errors made distinguishing between false positives, false negatives, true positives, and true negatives [156]. Breaking down the performance into these categories, enables a deeper analysis of model behavior, particularly in revealing biases towards certain classes or highlighting areas where the model may be confused.

As shown in Table 4.3 for the Xception model, the number of correct and incorrect predictions with count values are provided. The matrix quantifies the accuracy of the Xception model on the test dataset, where rows represent the actual categories, and columns represent the predicted categories. The matrix shows a high number of true positives (66.85%) and true negatives (32.10%), which indicates effective classifier performance. The low numbers of false negatives (0.71%) and false positives (0.34%) further demonstrate the model's high sensitivity and specificity in dis-

tinguishing between "interested" and "not-interested" responses. This visualization supports the evaluation of the predictive accuracy and error types made by the model in the context of viewer interest assessment.

Table 4.3: Xception Model Confusion Matrix

		Predicted Labels	
		Interested	**Not Interested**
Actual Labels	**Interested**	66.85% (True Positive)	0.34% (False Negative)
	Not Interested	0.71% (False Positive)	32.10% (True Negative)

Similarly, Table 4.4 provides the confusion matrix for the ResNet-50 model and it details the correct and incorrect predictions across each class. This matrix allows us to quantify the accuracy of the ResNet-50 model on the test dataset in the same way. The matrix shows a high percentage of true positives (63.75%) and true negatives (30.14%), indicating effective classification. However, there are slightly higher percentages of false negatives (2.68%) and false positives (3.43%) compared to the Xception model.

Table 4.4: ResNet-50 Model Confusion Matrix

		Predicted Labels	
		Interested	**Not Interested**
Actual Labels	**Interested**	63.75% (True Positive)	3.43% (False Negative)
	Not Interested	2.68% (False Positive)	30.14% (True Negative)

F1 Score

The F1 score provides a balanced measure of a model's performance, ensuring that both false positives and false negatives are taken into account [157]. This balanced approach makes the F1 score particularly valuable in applications where the cost of misclassification is high, and accuracy

alone could be misleading due to skewed class distributions. The F1 score (Equation 4.12) reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4.12}$$

Tables 4.5 and 4.6 provide the quantitative analysis of the performance of two CNN models, Xception and ResNet50, in differentiating between "interested" and "not interested." For the Xception model, the precision, recall, and F1 scores are all exceptionally high at 0.99 for the "interested" class and slightly lower (0.99 and 0.98) for the "not interested" class. Additionally, the overall accuracy of the Xception model is 0.99. This indicates that the model is highly accurate in identifying instances of interest with few false positives or false negatives. Similarly, the ResNet-50 model demonstrates strong performance with precision (0.96), recall (0.95), and F1 scores (0.95) for the "interested" class, and slightly lower metrics for the "not interested" class, resulting in an overall accuracy of 0.94, which, while strong, is lower than the Xception model. These metrics reflect the robustness and effectiveness of both models across the dataset, indicating their suitability for classifying participant interest.

Table 4.5: Xception Model Classification Report

Label	**Precision**	**Recall**	**F1**
Interested	0.99	0.99	0.99
Not-Interested	0.99	0.98	0.98
Xception Accuracy = 0.99			

Table 4.6: ResNet50 Model Classification Report

Label	**Precision**	**Recall**	**F1**
Interested	0.96	0.95	0.95
Not-Interested	0.90	0.92	0.91
ResNet-50 Accuracy = 0.94			

4.5.3 Evaluation of Misclassified Instances

Since the Xception model performed slightly better than ResNet-50, we selected the Xception model for further examination on instances where the model's predictions diverged from the actual states of interest or disinterest, as demonstrated in the provided images (Figure 4.11). This detailed examination is imperative not only for understanding the limitations inherent in the current model but also for identifying crucial areas that may benefit from further refinement and development.

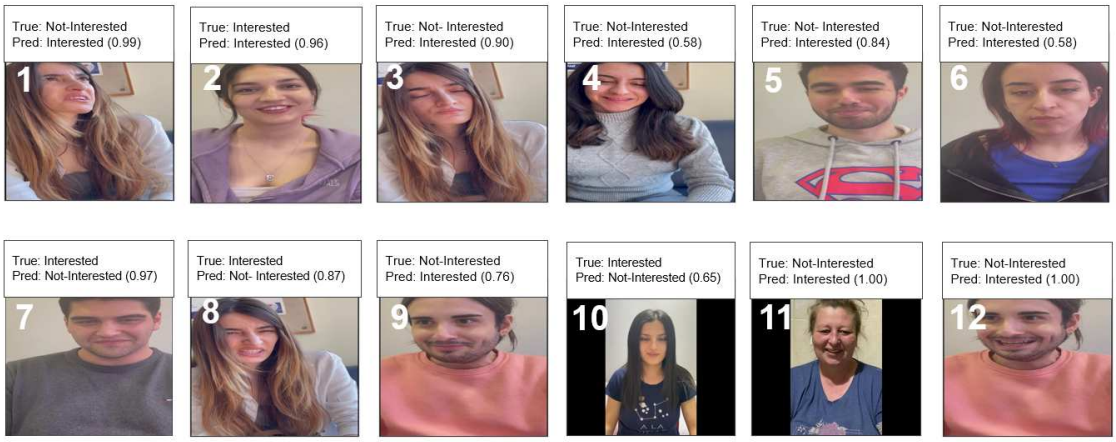


Figure 4.11: Analysis of Xception model inaccuracies in detecting emotional engagement.

Upon close inspection, several instances of misclassification were observed, highlighting a key challenge which is the model's handling of the subtle and complex nature of human facial expressions. Ambiguous expressions such as slight smirks (Figure 4.11 - Number 7) or neutral looks (Figure 4.11 - Number 6), which do not distinctly convey a single emotional state, were particularly problematic. An intriguing example includes an individual who appears to be laughing - a typically positive and engaged expression, yet the true label indicates that the person is not interested (Figure 4.11 - Numbers 9 and 12). This scenario confirms the complexity of interpreting expressions [158] that are highly susceptible to individual differences and contextual influences, aspects not currently accounted for by the model.

The factor contributing to these misclassifications is the variability in how individuals express emotions, which can be influenced by cultural backgrounds and personal idiosyncrasies [159].

Additionally, the quality and composition of the images, including factors such as lighting, angle, head rotation (Head Pose), and resolution, have a crucial impact on the model's performance [160]. These factors indicate a notable gap in the training dataset and suggest that the model's algorithm struggles to generalize across the diverse spectrum of emotional expressions and demographic backgrounds.

The implications of these misclassifications are significant, particularly in applications where precise emotion recognition is critical. To mitigate these issues, it is essential to enrich the training dataset with a broader variety of expressions and environmental conditions. Furthermore, the adoption of more sophisticated modeling techniques that can better adapt to the subtleties of human expressions may enhance the model's insight into genuine emotional states.

Thus, while the model demonstrates high precision and recall in general, the nuanced nature of human expressions necessitates further refinement. By addressing the highlighted misclassifications and integrating the suggested improvements, the model's applicability and reliability in real-world scenarios can be substantially increased, ensuring it meets the complex demands of accurately interpreting human emotions.

4.5.4 Correlation of Emotions with Interest Levels

To understand how participants' emotional responses (e.g., happy, angry, sad, etc) to advertisements correlate with their interest levels, further analysis was conducted. By examining the emotional responses, the aim is to identify which emotions are significant indicators of interest and which are associated with disinterest. To achieve this, the facial emotion recognition (FER) model [161] which uses Visual Geometry Group (VGG) architecture [140] was used to predict the probabilities of the seven basic emotions (anger, disgust, fear, happy, sad, surprise, neutral).

Figure 4.12 presents the pairwise relationships and distributions of the seven basic emotions between participants who are interested (blue points) and not interested (orange points) in the advertisements. Each plot compares one emotion against another, and the diagonal plots show the distribution of each emotion individually. The different colors in the Kernel Density Estimate

(KDE) plots in the lower triangle of the pairwise relationship figure represent varying levels of density where warmer colors (red or orange) represent higher density areas, while cooler colors (blue or purple) represent lower density areas. Also, the shape of the KDE plot indicates how the data points are distributed. For example, a circular shape (e.g., density plot between fear and happiness) indicates a more uniform distribution (weaker correlation), while an elongated shape (e.g., density plot between fear and neutral) suggests there is a stronger correlation between the two emotions.

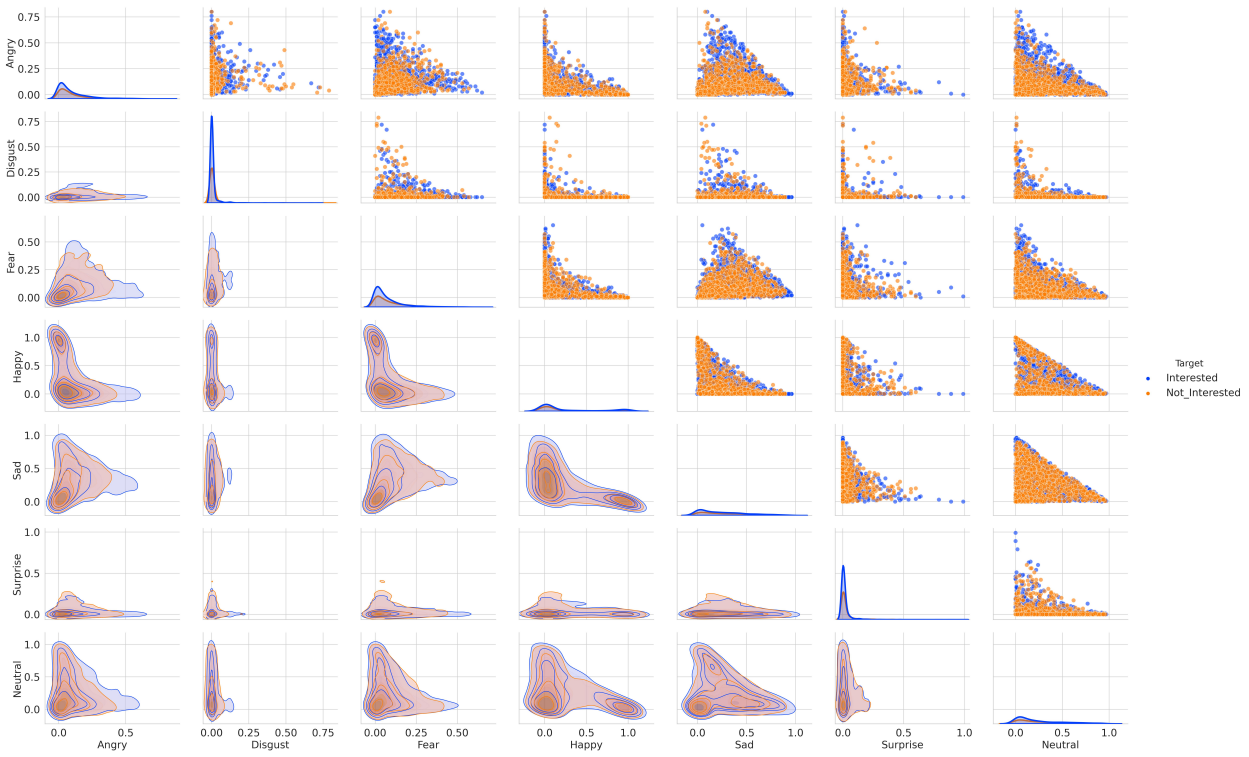


Figure 4.12: Pairwise distribution and correlation of emotional responses between interested and not-interested classes.

Additionally, Figure 4.13 shows the density plots for each emotion, highlighting the differences in distribution for the interested and not-interested classes by revealing distinct patterns in the emotional responses. These plots provide a clear view of how each emotion varies between the two groups, supporting the findings from the scatter plots in the pairwise Figure 4.12. From this figure, the inference can be made that happiness emerges as a significant indicator of interest, with higher

values observed in the interested group. In contrast, emotions such as disgust, fear, and surprise are more pronounced in the not-interested group which suggests that these negative emotions correlate with disinterest. Also, angry, sad, and neutral emotions show less clear differentiation between the two classes.

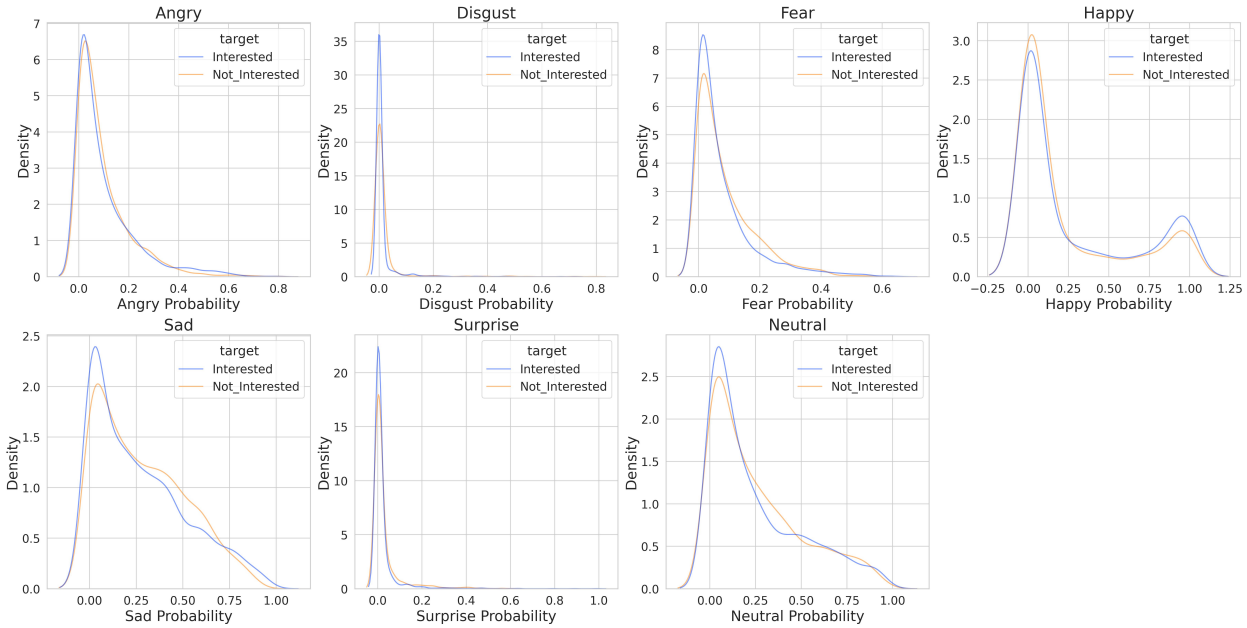


Figure 4.13: Density plots of emotional response probabilities for interested and not interested participants.

A series of t-tests were performed to compare the mean probabilities for each emotion between the interested versus not interested classes. The results, as summarized in Table 4.7, indicate significant differences in the mean probabilities for certain emotions between interested and not interested. Specifically, disgust, fear, happiness, and surprise showed statistically significant differences, with p-values below the threshold of α=0.05.

The analysis reveals that certain emotions, such as happiness, disgust, and fear, play a significant role in differentiating between the interested and not-interested classes. Higher levels of happiness are associated with interest, while higher levels of disgust and fear correlate with disinterest. This aligns with the hypothesis that positive emotions are indicative of consumer interest, while negative emotions suggest the opposite.

Table 4.7: T-Test Results Comparing Mean Emotion Probabilities for Interested and Not Interested Classes

Emotion	**Mean** *(Interested)*	**Mean** *(Not Interested)*	**Higher Mean**	**t-statistic**	**p-value**	**Significant**
Angry	0.10	0.10	Interested	0.92	0.355	No
Disgust	0.01	0.02	Not Interested	-2.49	0.013	Yes
Fear	0.08	0.09	Not Interested	-3.22	0.001	Yes
Happy	0.28	0.23	Interested	3.59	0.0003	Yes
Sad	0.27	0.28	Not Interested	-1.82	0.069	No
Surprise	0.02	0.03	Not Interested	-3.64	0.0003	Yes
Neutral	0.25	0.26	Not Interested	-0.96	0.339	No

The findings suggest that while FER models can offer detailed insights into emotional reactions, binary classifiers specially equipped with advanced architectures may be more practical for predicting consumer interest in a commercial context. The reduction in complexity and noise makes binary classifiers a robust choice for applications where the primary goal is to determine interest levels accurately.

4.6 Conclusions

This chapter highlights the significant impact of convolutional neural networks (CNNs) in enhancing digital marketing strategies through precise analysis of consumer emotional responses to advertisements. Employing advanced CNN architectures, specifically Xception and ResNet-50, our research successfully demonstrated the ability of these models to discern between “interested" and “not interested" consumer responses based on facial expression analysis. The findings reveal that emotions such as happiness significantly indicate consumer interest, while emotions like disgust and fear are more often associated with disinterest.

Our comparative analysis of the two CNN models highlighted the superior performance of the Xception model in terms of precision, recall, and the ability to generalize from training to unseen data. This was evidenced by its higher accuracy and finer understanding of nuanced emotional expressions compared to the ResNet-50 model. Furthermore, the statistical analysis reinforced the

practical implications of deploying facial emotion recognition technologies in real-time marketing scenarios to dynamically tailor content based on consumer emotional feedback.

The research also brought to light the complexities and challenges in interpreting subtle emotional cues, which vary widely among individuals and contexts. These findings emphasize the need for incorporating a diverse range of emotional expressions and demographic variables to enhance model robustness and accuracy.

In conclusion, the integration of CNNs into marketing analytics represents a transformative advancement, enabling businesses to not only predict consumer behavior with greater accuracy but also to engage in more personalized and effective marketing practices. As the integration of convolutional neural networks (CNNs) into neuromarketing and consumer behavior analysis continues to show promise, there is a vast landscape of potential research avenues that could further enhance the depth and applicability of this technology. One important consideration is the exploration of multimodal data sources. By integrating physiological signals such as heart rate and skin conductance with auditory and visual data, we may be able to achieve a more comprehensive understanding of consumer responses. Another critical area involves the cross-cultural validation of these models. Emotional expressions and their interpretations can vary significantly across different cultural contexts. It is crucial to validate and adapt CNN models across diverse demographic and cultural backgrounds to ensure their global applicability and sensitivity to cultural nuances in emotion recognition.

The development of real-time consumer feedback systems represents a transformative application of CNNs. Such systems could provide immediate feedback on consumer emotional responses during marketing campaigns, allowing marketers to adjust content dynamically. This capability would enable marketers to optimize engagement based on real-time emotional data, revolutionizing the way marketing strategies are implemented. Further, there is a need for advanced architectural innovations. New CNN architectures tailored specifically to the challenges of emotional recognition in marketing could significantly enhance model performance. Additionally, conducting longitudinal studies could provide valuable insights into how consumer responses to advertise-

ments evolve over time and across different contexts. Such studies could help in understanding long-term consumer behavior trends and the lasting impacts of marketing strategies.

Chapter 5

Aim 2: Addressing Dataset Limitations with Generative AI and Social Media Data

5.1 Aim 2 Summary

Generalizing deep learning models across diverse content types is a persistent challenge in domains like Facial Emotion Recognition (FER), where datasets often fail to reflect the wide range of emotional responses triggered by different stimuli. This chapter addresses the issue of content generalizability by comparing FER model performance between models trained on video data collected in a controlled laboratory environment, data extracted from a social media platform (YouTube), and synthetic data generated using Generative Adversarial Networks. The videos focus on facial reactions to advertisements, and the integration of these different data sources seeks to address underrepresented advertisement genres, emotional reactions, and individual diversity. Our FER models leverage convolutional neural network Xception architecture, which is fine-tuned using category-based sampling. This ensures training and validation data represent diverse advertisement categories while testing data includes novel content to evaluate generalizability rigorously. Precision-recall curves and ROC-AUC metrics are used to assess performance. Results indicate a 7% improvement in accuracy and a 12% increase in precision-recall AUC when combining real-world social media and synthetic data, demonstrating reduced overfitting and enhanced content generalizability. These findings highlight the effectiveness of integrating synthetic and real-world data to build FER systems that perform reliably across more diverse and representative content.

5.2 Introduction

The generalizability of deep learning models, particularly in the domain of Facial Emotion Recognition (FER), remains a significant challenge due to limited and non-diverse training datasets.

As a result, some artificial intelligence (AI) models may induce unintended biases, such as inaccurately classifying or recognizing types of people or contexts. While FER models offer numerous benefits across diverse domains, it is important to uphold ethical responsibilities in the development and deployment of these models [162].

Current and previous research has explored deep learning generalizability, yet many challenges still exist. Addressing these gaps, this chapter investigates the efficacy of integrating synthetic data generated through StyleGAN2 and real-world social media data from YouTube to enhance model generalizability. This chapter seeks to answer the critical research question of, “can FER model generalizability be improved using diverse data extracted from social media and generated through generative AI?" By systematically combining controlled-based data (Neurobiosense dataset), StyleGan2 Gen AI synthetic, and real-world YouTube data, this research aims to create FER systems that perform reliably across real-world scenarios.

5.2.1 Problem Statement

While previous studies have addressed various aspects of dataset diversity and model generalizability in FER, there remains a need for approaches that integrate synthetic and real-world data to overcome data constraints comprehensively. This chapter aims to fill this gap by demonstrating how synthetic data and social media data can be used to create more diverse and representative datasets for deep learning applications. In this study, we demonstrate FER in the context of interest/disinterest in digital advertisements. This paper leverages novel datasets to compare FER models trained using data from a controlled experiment, data augmented with synthetic images generated using Generative AI, and real-world data extracted from YouTube. Through the inclusion of emotional responses to broader content and supplementing underrepresented areas in emotion categories with synthetic data, this research seeks to enhance the generalizability of FER models. The following research questions are addressed in this chapter:

- **RQ 2.1:** Can FER model generalizability be improved using data extracted from social media (YouTube) and/or generated using AI (GANs), as compared to controlled data from a laboratory study?
- **RQ 2.2:** How can FER models trained on specific categories of advertisements be generalizable to new categories of advertisements?
- **RQ 2.3:** What are the ethical and practical considerations in using synthetic and real-world data for FER model training?

5.3 Data Description

In this chapter, we utilized three different data sources of human facial expressions: 1) data collected in the NeuroBioSense experiment [112], 2) data we extracted from YouTube, and 3) data we created using a generative AI platform. This data was used to train three separate Xception-based Facial Expression Recognition (FER) models. Then, the three models were compared for performance using a subset of the NeuroBioSense data reserved for validation and a subset of the NeuroBioSense reserved for testing.

These datasets were carefully designed to systematically assess the model's generalization capabilities across diverse participant demographics and various advertisement categories. The purpose of employing multiple datasets was to rigorously test the robustness and adaptability of the model in recognizing and interpreting facial expressions in different contexts.

5.3.1 NeuroBioSense (Baseline) Dataset

The first dataset, referred to as Baseline, is from the NeuroBioSense dataset [112] as used in the previous chapter for Research Aim 1, and represents data collected in a controlled laboratory-style environment. It consists of videos of the facial reactions of participants while they were exposed to three distinct advertisement categories: 1) car and technology, 2) food and market, and 3) cosmetics and fashion. Each participant was assigned to one advertisement category, ensuring that the reactions were specific to that category. There were 20 participants assigned to car and technol-

ogy, 20 to food and market, and 18 to cosmetics and fashion. Participants' facial expressions were captured in real-time and labeled based on their self-reported emotional responses, as interested or not-interested. This dataset includes a total of 58 participants (30 female, 23 male), ages 18 to 66 (mean = 27.4, SD = 11.3). The structured nature of the dataset provided a consistent and controlled environment, allowing for a reliable baseline in the evaluation of the FER model.

Recall, that the dataset includes a total of 1,045 video recordings capturing participants' facial expressions while watching the various advertisements. During the preprocessing phase, we systematically extracted frames from each of the 1,045 video recordings to capture consistent facial expressions across the viewing of advertisements. Specifically, we extracted 10 frames from each video at equal intervals, determined by dividing the total number of frames by 10. This method ensured that frames were uniformly sampled throughout the entire duration of each video, which provides a representative temporal cross-section of facial expressions and avoids random selection to minimize sampling bias and enhance the reliability of subsequent analyses. Thus, this resulted in a total of 10,450 images for our analysis. Specifically, there were 3,020 images of participant faces while watching the car and technology ads (2,510 labeled as interested, 83%); 3,370 images of participant faces while watching the food and market ads (1,860 labeled as interested, 55%); 4,060 images of participant faces while watching the cosmetics and fashion ads (2,650 labeled as interested, 65%).

The baseline data was divided into a train set, a validate set, and a test set. The images of the participants watching the food/market were used for training, the images for cosmetics/fashion were used for validation, and the car/technology was saved for model testing. This was intentionally done to ensure that model validation and testing occurred on novel sets of people and ad content than the models were trained on.

The baseline train data was further augmented with the additional data sources, described below, for the other two models. However, all three models were validated using the same 4,060 images and tested using the same 3,020 images.

5.3.2 NeuroBioSense and YouTube Combined (Baseline + YouTube) Dataset

To further diversify the dataset and enhance the model's ability to generalize, we expanded the NeuroBioSense data by incorporating a collection of YouTube reaction videos. These videos capture users' natural reactions to a variety of advertisements, providing a more dynamic and uncontrolled dataset compared to the original, which was collected in a more structured environment. The data that support the findings of this study are available from the author's GitHub [163].

The YouTube reaction dataset focuses on participants' real-time emotional responses to advertisements, with reactions categorized as either interested or not-interested. This addition introduced several complexities absent in the NeuroBioSense dataset, including greater diversity in participants. Furthermore, the types of advertisements covered were broader, introducing a wider range of products and scenarios, which allowed the model to become more robust in handling real-world conditions.

To collect this data, we took screenshots of videos available on YouTube. We found these videos by searching for "reaction videos to advertisements," with the inclusion criteria that the video contained participants that were verbal, at some point, about their interest/disinterest in the ad content. Additionally we sourced reaction footage from publicly available YouTube videos, we used featured titles such as: "Watching Ads That Are Unbelievably Funny," "Reviewing the Most Controversial Advertisements Ever," "Reacting to Hilariously Cringy Commercials," and "Adults React To Ads You Won't Believe Actually Aired! | REACT."

In total, we used 137 different YouTube videos of 137 different people (44 female, 93 male) and tagged the videos manually as interested or not-interested based on what they said in the video. For each of the 137 people, there were on average 14 images (SD = 3.44 images) of them included in the dataset. As a result, the YouTube dataset adds an additional 2,000 images to the baseline dataset. For these additional images, 1,000 were tagged as interested and 1,000 as not-interested.

This augmentation of the baseline data reflects an increase in diversity in both participant reactions and content exposure, helping to ensure that the model can generalize effectively across different user types and advertising contexts.

5.3.3 NeuroBioSense and Generative AI-Enhanced Synthetic Dataset (Baseline + StyleGAN2)

In the third dataset, we introduced synthetic data to further augment the baseline training data. This dataset was created using RunwayML, a generative AI platform that leverages advanced techniques like Style-based Generative Adversarial Networks (StyleGAN2) developed by NVIDIA, and designed for generating high-quality realistic synthetic images of human faces [53]. The primary purpose of this synthetic data was to introduce even more diversity in facial features, such as hairstyle, race, background, and clothing, while maintaining consistency in emotional expressions.

The synthetic data was created using the labeled (interested or not-interested) NeuroBioSense training data. For each video in the original NeuroBioSense training data, we randomly extracted one frame for use in generating the synthetic data. For each of these frames, RunwayML generated six images, resulting in an additional 2,022 images. This data was then incorporated into the training set to improve the model's ability to generalize across participants with unseen facial characteristics and features.

The use of generative AI allowed us to significantly increase the diversity of facial expressions and demographic variables without needing additional real-world data. This synthetic augmentation was instrumental in enhancing the model's ability to generalize across new faces, helping to overcome limitations encountered in conventional data augmentation techniques.

5.3.4 Dataset Combinations for Models

The three datasets described above were used to create the three conditions (Baseline, Baseline + YouTube, Baseline + StyleGAN2) for use in the FERs. Table 5.1 summarizes the number of images used in training, validating, and testing each model. Note, that the validate and test sets were the same across all three models; only the training sets differed. Additionally, the test data consisted of a subset of participants and advertisement categories that were not seen in any training data.

Table 5.1: Number of Images Used in Training, Validating, and Testing Each Model

Model	Train	Validate	Test	Total
Baseline	3,370	4,060	3,020	10,450
Baseline +YouTube	5,370	4,060	3,020	12,450
Baseline+ StyleGAN2	5,392	4,060	3,020	12,472

To ensure the robustness and generalizability of our models, we enhanced the diversity of our training datasets through the integration of social media and generative AI-generated data (Figure 5.1). Specifically, our datasets comprise participants across a broad spectrum of demographics, including gender, racial/ethnic diversity, and age groups. This approach aligns with recommendations from prior research emphasizing the importance of demographic diversity for developing FER systems that can accurately interpret facial expressions across various populations [42]. Integrating real-world data from social media platforms, such as YouTube, enables our models to capture context-rich emotional expressions that are often missing in controlled experimental setups [164]. Additionally, we used Generative Adversarial Networks (GANs) to supplement our dataset with synthetic images representing underrepresented groups, providing facial diversity that mirrors real-world conditions [165]. Compared to traditional FER datasets like CK+ and JAFFE, which often lack sufficient demographic representation [39, 40]. Also, to enhance the contextual and cultural diversity of our dataset, we included a variety of clothing styles, such as traditional, casual, and formal attire from different cultures. Clothing serves as a visual cue, providing context for emotion interpretation by reflecting cultural identity, seasonality, and situational factors. This diversity reduces model bias and prevents overfitting to specific patterns present in limited datasets [166].

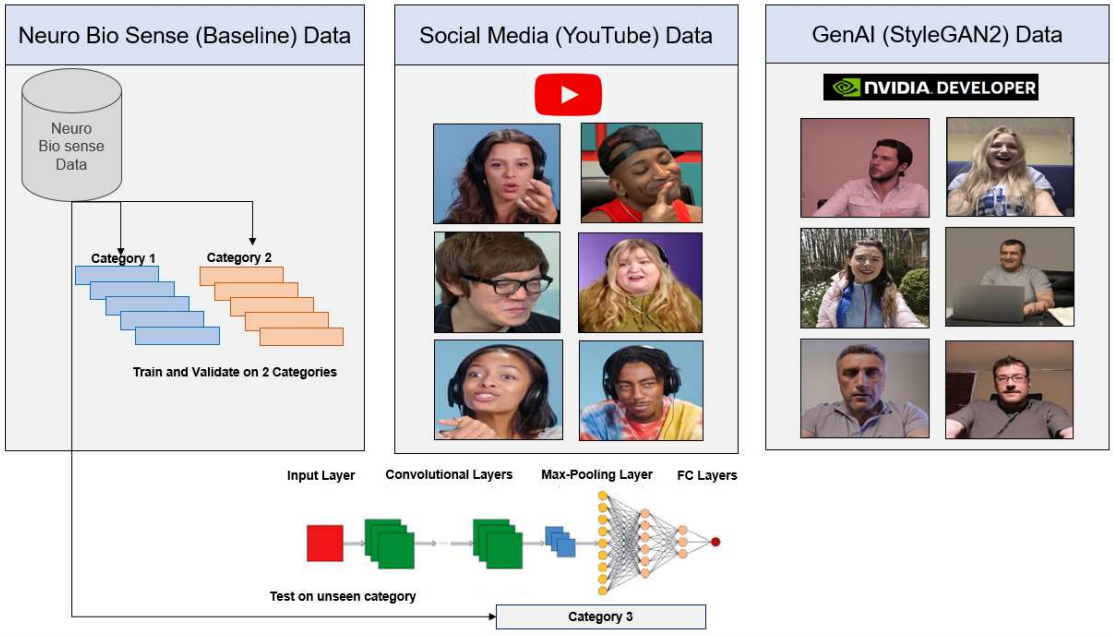


Figure 5.1: Overview of the training and validating pipeline integrating NeuroBioSense (Baseline), Social Media (YouTube), and Generative AI (StyleGAN2) datasets.

5.4 Theoretical Framework: StyleGAN2

The resolution and quality of images produced by generative methods, especially generative adversarial networks (GAN), are improving rapidly [167]. StyleGAN2 is a generative adversarial network (GAN) architecture that advances the capabilities of its predecessor, StyleGAN, by addressing specific limitations related to image quality and the presence of artifacts [168]. Developed by NVIDIA researchers [53], StyleGAN2 introduces significant architectural modifications and new techniques to enhance the fidelity, realism, and controllability of synthesized images.

StyleGAN2 maintains specific facial details like expressions while altering other elements such as the background. This is achieved through its hierarchical architecture that modulates different aspects of the image at various layers [66]. The model employs a mapping network that transforms an input latent vector into intermediate latent codes. These codes influence the generator's layers differently, with early layers controlling high-level attributes like facial structure and expressions, and later layers affecting fine-grained details like texture and background.

By keeping the latent codes consistent for early layers, StyleGAN2 preserves facial expressions. At the same time, it varies the codes for later layers, allowing background changes. This approach enables StyleGAN2 to disentangle features effectively [169]. As a result, it can keep details, like a consistent laugh on the face, while other aspects evolve independently. This is achieved through the model's style-based synthesis and progressive refinement mechanisms within the generator network [170].

At the core of StyleGAN2 is the concept of style-based synthesis [171], where a mapping network transforms a latent vector $\mathbf{z} \in \mathcal{Z}$ into an intermediate latent space $\mathbf{w} \in \mathcal{W}$. This intermediate vector modulates the generator to produce images with disentangled attributes [172].

In the original StyleGAN, the generator employed adaptive instance normalization (AdaIN) layers to inject style information [173]. However, this approach led to characteristic artifacts, such as droplet-shaped distortions, due to inherent biases introduced by normalization operations [174]. To eliminate these artifacts, StyleGAN2 replaces AdaIN with a novel mechanism called weight demodulation [53].

Instead of normalizing the activations, StyleGAN2 modulates the weights of the convolutional layers directly based on the style vector. Mathematically, the modulation and demodulation processes are defined as follows:

The convolutional weights w_{ijk} are modulated by the style coefficients s_i:

$$\hat{w}_{ijk} = s_i \cdot w_{ijk}, \tag{5.1}$$

where i indexes the input feature maps, j the output feature maps, and k the spatial kernel positions.

To prevent the amplification of certain features and maintain consistent signal magnitudes, the modulated weights are demodulated:

$$\tilde{w}_{ijk} = \frac{\hat{w}_{ijk}}{\sqrt{\sum_i (s_i \cdot w_{ijk})^2 + \epsilon}}, \tag{5.2}$$

where ϵ is a small constant added for numerical stability.

The demodulation step effectively normalizes the weights, ensuring that the style modulation does not introduce undesired biases or artifacts [175]. This approach maintains the statistical properties of the feature maps without relying on explicit normalization layers.

Another critical innovation in StyleGAN2 is the introduction of path length regularization [53]. This technique encourages the generator to produce images that respond smoothly and predictably to changes in the latent space $\mathcal{W}$. The regularization term R_{pl} is defined as:

$$R_{\text{pl}} = \mathbb{E}_{\mathbf{w},\mathbf{y}\sim\mathcal{N}(0,I)} \left(\left\| \mathbf{y}^\top \frac{\partial G(\mathbf{w})}{\partial \mathbf{w}} \right\|_2 - a \right)^2, \tag{5.3}$$

where $G(\mathbf{w})$ is the generator output (image) given the latent vector $\mathbf{w}$, $\mathbf{y}$ is a random vector sampled from a standard normal distribution and a is an exponential moving average of the path lengths to stabilize training.

This regularization penalizes deviations from the expected rate of change in the images with respect to the latent vectors, promoting a more linear and interpretable latent space.

StyleGAN2 also refines the mapping network and the hierarchical application of styles. By adjusting where and how the styles are applied within the generator, the architecture achieves a better separation of high-level attributes (such as pose and identity in face generation) from fine-grained details (like texture and color). The mapping network $f : \mathcal{Z} \to \mathcal{W}$ is designed to increase the expressiveness of the latent space $\mathcal{W}$, which enables more nuanced control over the generated images.

The discriminator in StyleGAN2 is augmented with techniques like lazy regularization [53], where computationally intensive regularization terms are applied less frequently to reduce training overhead without sacrificing performance. The discriminator loss includes an R1 regularization term, which penalizes the gradient norm of the discriminator's output with respect to the input images:

$$R_1 = \frac{\gamma}{2} \mathbb{E}_{\mathbf{x}\sim P_{\text{data}}} \left(\|\nabla_{\mathbf{x}} D(\mathbf{x})\|_2^2 \right), \tag{5.4}$$

where:

- $D(\mathbf{x})$ is the discriminator's output given real image $\mathbf{x}$.
- γ is a regularization coefficient.
- P_{data} is the distribution of real images.

The culmination of these architectural and methodological advancements allows StyleGAN2 to generate images with unprecedented quality. The elimination of normalization biases, improved modulation techniques, and enhanced regularization contribute to the synthesis of images that are both highly realistic and controllable. The generator can produce high-resolution images (e.g., 1024×1024 pixels) that exhibit fine details.

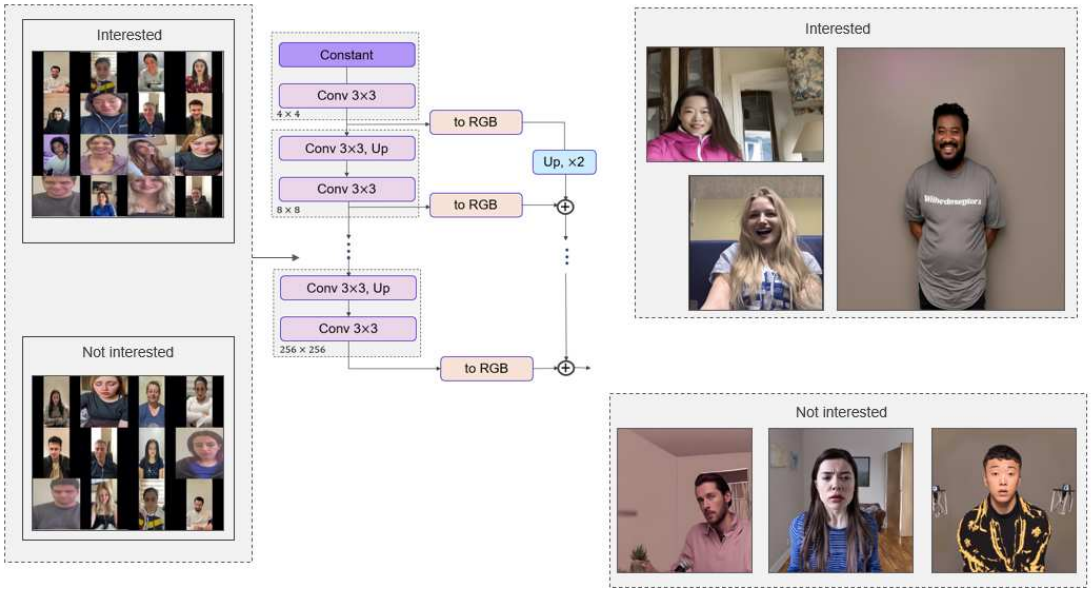


Figure 5.2: Synthetic image generation using a StyleGAN model based on 'interested' and 'not interested' image categories.

5.5 Model Training and Optimization Strategy

There were three separate models trained using the data combinations. The inclusion of diverse stimuli, such as advertisements across various categories, enhances the model's ability to generalize across different content types.

We formulated the FER task as a binary classification problem, distinguishing between expressions of interest and non-interest. The model optimization involves minimizing the binary cross-entropy loss function, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{5.5}$$

where N is the number of samples, $y_i \in 0, 1$ is the true label, and p_i is the predicted probability that the i-th sample belongs to the positive class.

We employed the Adam optimizer [176], which adapts the learning rate for each parameter using estimates of the first and second moments of the gradients. The parameter updates are computed as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{5.6}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{5.7}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{5.8}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{5.9}$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{5.10}$$

where g_t is the gradient at time step t, m_t and v_t are the exponentially decaying averages of past gradients and squared gradients, β_1 and β_2 are decay rates, α is the learning rate, and ϵ is a small constant to prevent division by zero.

The learning rate was initialized at $\alpha = 1 \times 10^{-4}$ and decayed exponentially to ensure convergence and prevent overfitting. Training was conducted over 300 epochs with a batch size of 16. After finalizing the optimization process and model parameters, we utilized Google Cloud's e2-highmem-16 instance equipped with an NVIDIA Tesla P100 GPU to handle the computational demands of training the Convolutional Neural Networks.

5.5.1 Data Augmentation and Regularization

To enhance model robustness and generalization, we applied various data augmentation techniques to the training data [177]. First, we implemented geometric transformations, including random rotations with angles θ in the range $[-15°, 15°]$, horizontal flips, and random cropping to simulate different viewing angles and facial orientations. Additionally, photometric adjustments such as random changes in brightness, contrast, and saturation were applied to mimic varying lighting conditions. To account for sensor noise, Gaussian noise with zero mean and small variance was added to the images [177].

Regularization techniques were also employed to prevent overfitting. We applied dropout regularization with a dropout rate of 0.5 to the fully connected layers, randomly deactivating neurons during training [178]. Furthermore, weight decay, also known as L2 regularization, was incorporated by adding a penalty term $\lambda\|\theta\|_2^2$ to the loss function, where λ is the regularization parameter and θ represents the model weights [179]. This penalty discourages the network from assigning excessively large weights, promoting simpler models that generalize better.

5.6 Experimental Design

The experimental design of this study aims to assess the generalizability of FER models by using data from diverse sources, including both real-world and synthetic datasets. To ensure a robust and fair evaluation of model performance, a category-based data-splitting strategy was adopted. This approach ensures that the training, validation, and test subsets are separated based on adver-

tisement categories, allowing the models to be trained and validated on specific categories while being tested on entirely new categories.

The rationale for using category-based data splitting lies in the need to evaluate the model's ability to generalize beyond the categories it has been trained on. This approach mimics real-world scenarios where FER systems encounter entirely new content types. By training on one set of categories and testing on different, unseen categories, the design measures the model's capacity to adapt to novel contexts, a critical aspect of enhancing model robustness.

The category-based splitting approach offers several advantages for evaluating the FER models. First, it provides a realistic assessment of the model's ability to handle new advertisement categories, reflecting real-world deployment scenarios where models often encounter unfamiliar stimuli. Second, this strategy reduces potential biases that could arise from overlapping content between training, validation, and test sets, ensuring that the model's performance reflects true generalizability rather than overfitting to specific patterns.

5.7 FER Model Framework

5.7.1 Xception

In this study, we employ the Xception model, a convolutional neural network (CNN) architecture that has demonstrated superior performance in previous research by Alipour et al. [4]. The Xception model builds upon the Inception architecture [180] by replacing the standard Inception modules with depthwise separable convolutions, thereby enhancing both computational efficiency and model accuracy. Our selection of the Xception architecture is motivated by its ability to optimize resource utilization while achieving high performance in large-scale visual recognition tasks [181], making it particularly suitable for Facial Emotion Recognition in diverse content environments.

5.7.2 Xception Model Architecture

The term Xception stands for Extreme Inception [128], reflecting its evolution from the original Inception model. The key innovation in Xception is the use of depthwise separable convolutions, which decompose a standard convolution into a depthwise convolution and a pointwise convolution. This decomposition allows for a more efficient representation of the convolutional operation, reducing the number of parameters and computational costs without sacrificing model capacity.

A standard convolutional operation combines spatial filtering and channel mixing in a single step. Mathematically, a standard convolution for an input tensor $\mathbf{X} \in \mathbb{R}^{h \times w \times c_{\text{in}}}$ with c_{in} input channels, applying c_{out} filters of size $k \times k$, produces an output tensor $\mathbf{Y} \in \mathbb{R}^{h' \times w' \times c_{\text{out}}}$, computed as:

$$\mathbf{Y}i, j, k = \sum u = 1^k \sum_{v=1}^{k} \sum_{c=1}^{c_{\text{in}}} \mathbf{W}u, v, c, k \cdot \mathbf{X}i + u - 1, j + v - 1, c \tag{5.11}$$

where $\mathbf{W} \in \mathbb{R}^{k \times k \times c_{\text{in}} \times c_{\text{out}}}$ represents the convolutional filters.

In contrast, a depthwise separable convolution splits this operation into two separate layers of depthwise and pointwise.

Depthwise convolution which applies a single convolutional filter per input channel independently.

$$\mathbf{Z}_{i,j,c} = \sum_{u=1}^{k} \sum_{v=1}^{k} \mathbf{K}_{u,v,c} \cdot \mathbf{X}_{i+u-1,j+v-1,c} \tag{5.12}$$

where $\mathbf{K} \in \mathbb{R}^{k \times k \times c_{\text{in}}}$ is the depthwise convolutional filter.

Also, the pointwise convolution which applies a 1×1 convolution to combine the outputs of the depthwise convolution across channels.

$$\mathbf{Y}_{i,j,k} = \sum_{c=1}^{c_{\text{in}}} \mathbf{P}_{c,k} \cdot \mathbf{Z}_{i,j,c} \tag{5.13}$$

where $\mathbf{P} \in \mathbb{R}^{c_{\text{in}} \times c_{\text{out}}}$ represents the pointwise convolutional filters.

This separation reduces the computational complexity significantly lowering the number of parameters and operations.

The Xception architecture is organized into three main parts of entry flow, middle flow, and exit flow. The entry flow is responsible for capturing low-level features and reducing the spatial dimensions of the input data. It typically uses convolutional layers followed by max-pooling operations to down-sample the data.

Next, the middle flow is made up of several identical modules (repeated eight times) designed to learn increasingly complex features. These modules contain depthwise separable convolution layers, which are more efficient than standard convolutions, and they also use residual connections to help preserve information as it passes through the layers.

Finally, the exit flow focuses on extracting high-level features and preparing the data for classification. This phase also uses depthwise separable convolutions, followed by global average pooling to reduce the dimensions further, and a fully connected layer for producing the final output suitable for classification tasks.

Residual connections are incorporated to mitigate the vanishing gradient problem and facilitate the training of deeper networks [182]. The overall architecture enables the model to learn rich feature representations essential for distinguishing subtle facial expressions in FER tasks.

5.7.3 Implementation Details

The model was implemented using the Keras library with a TensorFlow backend. The top layers were replaced with a global average pooling layer followed by a fully connected layer with a sigmoid activation function for binary classification and early stopping was implemented based on the validation loss to prevent overfitting.

5.7.4 Metrics for Evaluating Model Performance

To assess the effectiveness of our models in recognizing facial expressions and determining interest or non-interest, we employ two primary metrics: loss and accuracy [183]. These metrics offer a detailed understanding of the models' learning progress, capacity to generalize to novel

data, and overall robustness [184]. The combination of these metrics provides a balanced view of model performance across diverse datasets, highlighting how well the models have learned and adapted effectively to varying data patterns.

Loss Function

The loss function provides a scalar value that represents the cost associated with the network's predictions [185]. During training, the objective is to find the set of network parameters (weights and biases) that minimize this loss. This process is carried out using optimization algorithms; Adaptive Moment Estimation and stochastic gradient descent and its variants, which rely on the gradient of the loss function with respect to the network parameters [125].

Accuracy

Accuracy quantifies the proportion of correct predictions made by the model out of all predictions which provides a straightforward measure of how well the model generalizes to unseen data. Mathematically, for a dataset with n samples, accuracy A is calculated as:

$$A = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i = \hat{y}_i)$$

where:

- y_i is the true label of the i-th sample.
- $\hat{y}_i$ is the predicted label of the i-th sample.
- $\mathbb{I}$ is the indicator function, which returns 1 if the argument is true and 0 otherwise.

Accuracy serves as a primary metric during both the training and evaluation phases of convolutional neural network models. It provides an immediate sense of the model's ability to correctly classify input data. High accuracy indicates effective learning of the underlying patterns in the data, while low accuracy suggests the need for model refinement.

5.8 Results and Discussion

This section presents a comparative analysis of three models trained on distinct datasets: Baseline, Baseline + YouTube, and Baseline + StyleGAN2. The models are evaluated based on key performance metrics, including loss, accuracy, precision-recall curves, and ROC curves, providing discernment into their robustness, generalizability, and adaptability to diverse data sets.

5.8.1 Loss

Figure 5.3 illustrates the progression of training and validation losses across 300 epochs for each of the models to provide information about their learning dynamics and generalization capabilities. The Baseline + YouTube model (green line) demonstrates the most consistent validation loss, indicative of its capacity to generalize effectively to unseen data. Its training loss also remains low and aligns closely with the validation loss which shows that the model effectively learns from the diverse, real-world data without overfitting. The Baseline + StyleGAN2 model (blue line) initially displays a higher loss which suggests challenges in adapting to synthetic data, but ultimately converges. It highlights the potential of synthetic augmentation to effectively complement real-world data and eventually enhance model performance. Conversely, the Baseline model (red line) exhibits a rapid initial reduction in training loss during initial epochs that reflects its capacity to quickly adapt to the limited dataset, yet suffers from elevated validation loss that signifies limited generalization due to the constraints of its dataset. The persistent gap between its training and validation losses suggests overfitting, resulting from the constrained diversity of the training data, which limits the model's capacity to handle varied, real-world expressions.

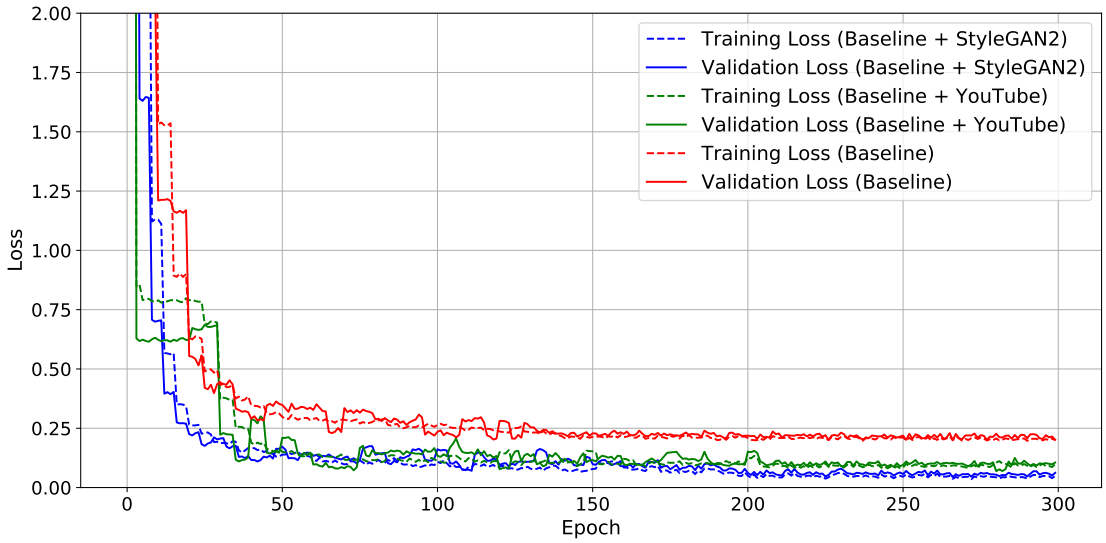


Figure 5.3: Comparison of training and validation losses across the three datasets over 300 epochs.

5.8.2 Accuracy

During training, accuracy is also monitored on both training and validation datasets to assess the model's learning progress and generalization ability. Figure 5.4 illustrates the trends in training and validation accuracy over 300 epochs for each model and it provides insights into their learning efficiency and generalization capabilities. The training accuracy indicates how well the models learn from their respective datasets, while validation accuracy measures how effectively the models generalize to unseen data.

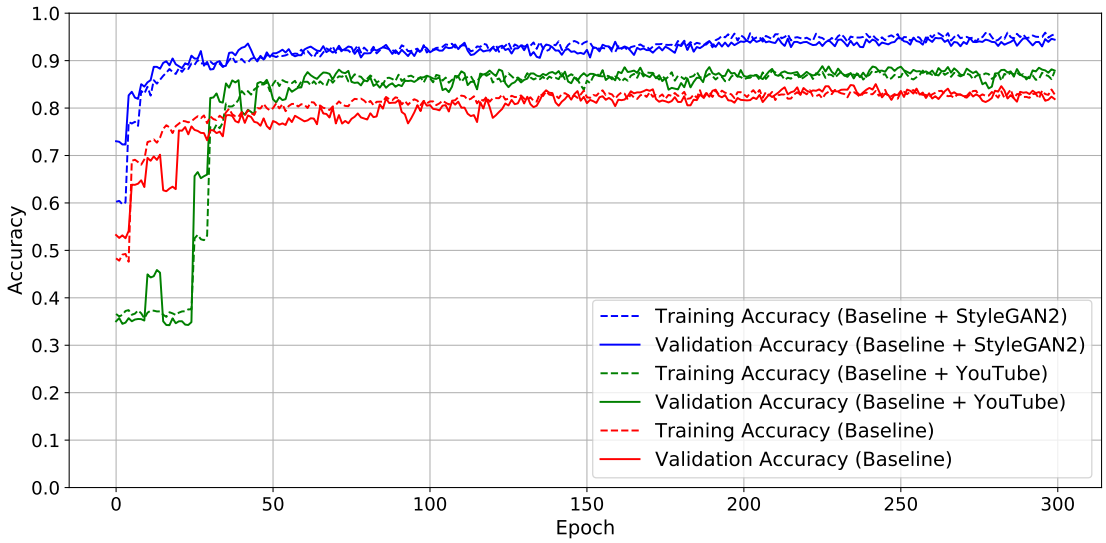


Figure 5.4: Comparison of training and validation accuracies across three datasets over 300 epochs.

The Baseline + YouTube model, represented in green, demonstrates a steady increase in training accuracy over the epochs that shows effective learning from the diverse, real-world data. Notably, the training accuracy aligns closely with validation accuracy which shows the model avoids overfitting and maintains robust generalization capabilities. This alignment can be attributed to the inclusion of varied expressions and contexts from the YouTube dataset, which exposes the model to a wide range of real-world patterns during training and enables it to handle unseen data effectively.

In contrast, the Baseline + StyleGAN2 model, depicted in blue, achieves rapid improvements in training accuracy as it reaches high levels early in the training process. This indicates that the model quickly adapts to the synthetic patterns present in the data. While this suggests effective utilization of synthetic augmentation, it also raises concerns about potential overfitting, as evidenced by a noticeable gap between training and validation accuracy. The validation accuracy, although improving over time, remains lower than training accuracy, suggesting that the synthetic data introduces patterns that are less representative of real-world variability. Consequently, the model's ability to generalize effectively is limited by the artificial structure inherent to synthetic data, which may not capture the complexity of natural expressions as accurately as real-world data does.

The Baseline model, shown in red, displays rapid gains in training accuracy, reflecting its quick adaptation to the limited patterns of the constrained dataset. However, this strong performance during training does not translate well to the validation set, where the accuracy remains significantly lower. The persistent gap between training and validation accuracy highlights the model's limited capacity for generalization, which is likely due to the homogeneity and restricted diversity of the dataset. As a result, the model tends to overfit the specific patterns present in the training data, failing to adapt to varied, unseen expressions.

5.8.3 Precision-Recall and Receiver Operating Characteristic Curve

Beyond the evaluation of loss and accuracy, the models' performance is further assessed using the Precision-Recall (PR) curve and the Receiver Operating Characteristic (ROC) curve. These metrics offer a more nuanced and detailed perspective, particularly in contexts where class imbalance may render accuracy alone an insufficient measure of model performance.

The Precision-Recall curve (Figure 5.5 - left) measures the trade-off between precision and recall across various decision thresholds. Precision, or positive predictive value, reflects the proportion of correctly identified positive cases among all predicted positives, while recall, or true positive rate, denotes the proportion of correctly identified positive cases among all actual positives. This metric is particularly informative in contexts where class imbalances are present, as it helps highlight the model's capability to minimize false positives while maintaining high recall. As depicted in Figure 5.5 - left, the Baseline + StyleGAN2 model achieves the highest area under the PR curve (AUC = 0.94), indicating a superior balance between precision and recall. The Baseline + YouTube model follows closely with an AUC of 0.89, suggesting strong generalizability to real-world data. In contrast, the Baseline model, with an AUC of 0.82, demonstrates weaker performance, suggesting challenges in maintaining both precision and recall due to limited data diversity.

The Receiver Operating Characteristic (ROC) curve (Figure 5.5 - right) provides another perspective by plotting the true positive rate against the false positive rate across varying thresholds.

The Area Under the Curve (AUC) serves as a summary metric, reflecting the model's overall ability to distinguish between positive and negative classes. In Figure 5.5 - right, the Baseline + StyleGAN2 model again achieves the highest ROC-AUC at 0.94 which indicates a strong discriminative capacity. The Baseline + YouTube model, with an AUC of 0.88, also exhibits robust performance which confirms its effectiveness in generalizing to diverse real-world data. The Baseline model, however, achieves a lower AUC of 0.82, highlighting its limited ability to effectively separate classes, which can be attributed to the constraints of its dataset.

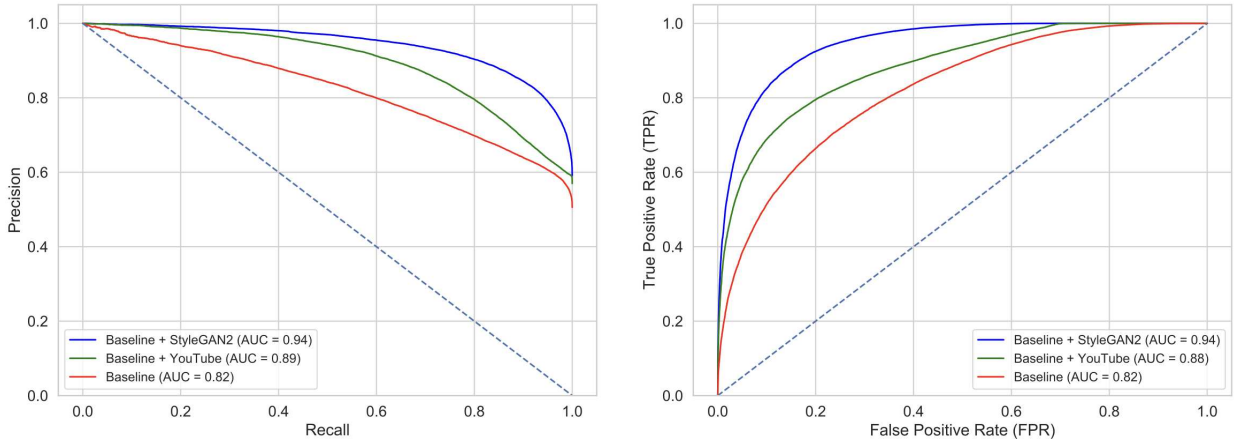


Figure 5.5: PR (left) and ROC (right) curves illustrating models' performance and balance across different datasets.

Additionally, Table 5.2 highlights significant percentage improvements in the Precision-Recall AUC and ROC-AUC metrics following the integration of diverse data sources into the Facial Emotion Recognition (FER) models. Initially, the Baseline model achieved a Precision-Recall AUC of 0.82. With the inclusion of real-world YouTube data (Baseline + YouTube), this metric rose to 0.89, representing an approximate 8.5% improvement over the baseline. This gain indicates a more balanced model performance, with better detection of true positives and reduced false positives. When synthetic data from StyleGAN2 (Baseline + StyleGAN2) was added, the Precision-Recall AUC increased further to 0.94, reflecting a 14.6% improvement over the baseline. This notable enhancement suggests that the synthetic data effectively addressed the variability in facial expressions, filling gaps left by the real-world dataset. Similarly, the ROC-AUC for the Baseline model

was initially 0.82, which increased to 0.88 with YouTube data, showing a 7.3% improvement. This indicates stronger discriminative power, as the model became more adept at distinguishing between positive and negative classes across a broader spectrum of emotional expressions. With the integration of synthetic data in the Baseline + StyleGAN2 model, the ROC-AUC reached 0.94, marking a 14.6% improvement.

Table 5.2: Performance Improvements for Precision-Recall AUC and ROC-AUC Across Models

Model/Dataset	**PR AUC**	**ROC AUC**	**Improvement from Baseline**
Baseline	0.82	0.82	–
Baseline + YouTube	0.89	0.88	+0.07 (PR), +0.06 (ROC)
Baseline + StyleGAN2	0.94	0.94	+0.12 (PR), +0.12 (ROC)

Lastly, to further understand the performance differences among the three models, we conducted an Analysis of Variance (ANOVA) on the AUC scores obtained from both the ROC and Precision-Recall curves. The ANOVA test aims to assess whether the mean differences in AUC scores across the models are statistically significant.

The ANOVA results revealed a statistically significant difference in the mean AUC scores across the three models ($F(2, 297) = 24.38, p < 0.001$). This indicates that the inclusion of YouTube and synthetic data led to significant improvements in model generalizability, as reflected in the AUC scores.

5.9 Ethical Considerations and Solutions

The integration of both real-world and synthetic data in Facial Emotion Recognition (FER) models holds significant potential for enhancing model performance and generalizability. However, it also raises complex ethical issues related to bias, fairness, and privacy [186]. This section examines these ethical challenges, discussing potential biases in FER models, privacy concerns arising from social media data usage [187], and proposing solutions to ensure the responsible de-

velopment and deployment of FER systems. By addressing these ethical considerations, the study aims to contribute to the creation of more inclusive and fair FER models [188].

5.9.1 Biases in Facial Emotion Recognition (FER) Models

FER models have gained significant traction across domains such as marketing, mental health assessment, and human-computer interaction [160]. Despite their potential, these models are susceptible to biases arising from the limited diversity of training datasets, which can result in skewed interpretations of facial expressions [189]. A common issue is the overrepresentation or underrepresentation of specific demographic groups, such as certain races, age groups, or genders, leading to model inaccuracies when applied to broader populations [26]. While this study aims to address these biases by integrating synthetic data, it is critical to observe that synthetic data generation itself may replicate and even amplify biases inherent in the original datasets [190]. For instance, GANs, which are used for creating synthetic data, can inadvertently perpetuate imbalances [191], especially if the initial data contains disproportionate samples from specific groups [192]. This tendency may affect the fairness and accuracy of FER models when deployed in real-world applications, making it imperative to examine and mitigate potential biases in both real-world and synthetic data sources [193].

5.9.2 Privacy Concerns in Social Media Data

The use of social media data to enhance FER models introduces complex ethical challenges, particularly concerning privacy, consent, and data protection [28]. Social media platforms offer a vast pool of user-generated content that can provide valuable diversity in facial expressions. However, this diversity comes with ethical dilemmas related to the rights of users whose content is used without explicit consent [194]. While social media data presents opportunities for improving model generalizability, it also poses risks of violating user privacy, as individuals may be unaware of their content being utilized for research or commercial purposes [195]. Additionally, the unstructured nature of social media data, which often includes personal information, raises ethical concerns about the extent to which data can be anonymized without compromising its utility for

training models [196]. Therefore, it is crucial to reflect on the ethical implications of data use, particularly in balancing the benefits of model enhancement with the rights of individuals to privacy and data protection.

5.9.3 Privacy Concerns in Synthetic Data

While synthetic data generation offers a promising solution for augmenting datasets [197] and alleviating privacy concerns, it also presents unique ethical and privacy challenges [198]. Synthetic data, particularly from Generative Adversarial Networks, can inadvertently reproduce identifiable patterns from original datasets [199], risking privacy and potentially perpetuating inherent biases. This issue is especially prominent with smaller datasets, where GANs may overfit on particular features, resulting in synthetic data that resembles real individuals with poor quality [200]. By addressing these privacy challenges, researchers can leverage synthetic data's benefits.

5.9.4 Addressing Ethical Issues and Promoting Fairness

To foster ethical integrity in FER model development, it is essential to establish robust measures that address bias mitigation and privacy protection [201]. First, conducting comprehensive data audits can play a vital role in identifying and correcting imbalances in demographic representation [202]. By analyzing demographic distributions in both real-world and synthetic datasets, researchers can ensure that training data aligns more closely with the intended deployment population [193]. This process should be coupled with the adoption of fairness-aware generative models, such as FairGAN [203] or DebiasGAN [204], which aim to produce synthetic data that better represent underrepresented groups without exaggerating biases found in the original data.

In addition to addressing biases in training data, it is equally important to consider the ethical implications of using social media and synthetic AI-generated data [205]. One solution is to ensure transparency in data collection processes. This involves using datasets that are licensed for research, prioritizing those where contributors have provided explicit consent for their data to be used [206]. Where consent is not feasible, researchers should employ anonymization techniques that protect individual identities while preserving the diversity of expressions needed for FER

model training [207]. Establishing clear protocols for obtaining permission and using ethically sourced datasets is crucial in maintaining the integrity of FER research.

Furthermore, integrating fairness metrics during model training can enhance the equity of FER models. Metrics such as Equalized Odds [208] and Demographic Parity [209] can help evaluate whether the model performs consistently across demographic groups, thus promoting fairness and minimizing the risk of discriminatory outcomes. It is imperative to implement these metrics throughout the model development process, from training to validation, ensuring that the model's generalization capabilities extend fairly across different demographic contexts.

Lastly, ethical compliance should extend beyond technical adjustments. Adherence to global data protection regulations, such as the General Data Protection Regulation (GDPR) [210] and the California Consumer Privacy Act (CCPA) [211], must guide all aspects of data handling. Researchers should collaborate with legal experts and ethics review boards to ensure compliance with these regulations, thereby aligning FER model development with legal standards of privacy protection [212]

5.9.5 Future Ethical Directions in FER Research

Looking ahead, it is essential to adopt more inclusive data collection strategies that are rooted in explicit consent and ethical sourcing. The implementation of differential privacy techniques holds promise for safeguarding individual privacy while allowing models to learn from data effectively. Moreover, advancing explainable AI (XAI) techniques in FER could enhance transparency which enables users to understand the basis of emotion predictions and identify potential biases in real-time [213]. Future research should prioritize these approaches, aiming to establish FER models that are both technologically robust and ethically sound.

5.10 Conclusions

In this chapter, we set out to improve the generalizability of an FER convolutional neural network Xception model by integrating diverse data sources. We combined real-world data from YouTube, controlled experimental data, and synthetic images generated by StyleGAN2.

The results demonstrate that integrating these diverse datasets significantly improves model performance. The models trained with the combined data sources consistently outperformed those trained on individual datasets, showing better accuracy and reduced overfitting. The addition of real-world data from YouTube enriched the models with more natural variations of facial expressions, closely reflecting real-world dynamics. Meanwhile, the synthetic data generated by StyleGAN2 filled critical gaps, introducing underrepresented facial expressions and enhancing the diversity of the training dataset.

These findings highlight the importance of data diversity in achieving model generalizability. However, this process also highlighted several ethical considerations. The use of social media data raises privacy concerns, particularly regarding consent and data protection, while synthetic data carries the risk of reinforcing existing biases if not properly managed. Thus, the study not only contributes to improving FER models technically but also emphasizes the need for responsible AI practices that balance innovation with ethical integrity.

In conclusion, while the integration of real-world and synthetic data represents a promising strategy for enhancing FER models, it also necessitates continuous refinement to ensure both technical performance and ethical compliance. Future research should focus on further improving the realism of synthetic data and exploring ways to ensure fairness and transparency in FER systems. This work thus lays the groundwork for developing FER models that are not only technically advanced but also ethically robust and capable of performing reliably across diverse, real-world scenarios.

Chapter 6

Aim 3: Effects of VR Design Complexity and Exposure Sequencing on Engagement

6.1 Aim 3 Summary

This chapter investigates how complexity and exposure sequencing in virtual reality (VR) retail environments influence consumer engagement and purchase intent. Participants (N = 55) interacted with two desktop VR settings: Simple and Detailed. Employing multi-modal analysis combining causal evaluation of pre- and post-exposure data, we assess how these environments influence user experiences. Results indicate Detailed VR significantly enhances participants' sense of immersion, engagement, and purchase intent compared to Simple VR. This enhanced immersion does not correspond to increased perceived cognitive load, suggesting that added complexity enriches user experience without adverse effects. The sequence of VR exposure also plays a critical role. Participants who began with Simple and then Detailed VR report significantly higher engagement in the latter; suggesting that transitioning from a simpler to a more complex environment amplifies user engagement through comparative effects. This research contributes to immersive marketing literature by isolating the effects of VR design complexity and exposure order on consumer behavior.

6.2 Introduction

This research aim addresses critical gaps in the literature regarding how virtual reality environments influence consumer behavior particularly engagement and purchase intent. While prior research has highlighted VR's potential to enhance user experiences through immersive and interactive features there is limited understanding of the specific design elements that drive these effects. Moreover, existing studies often overlook the role of individual differences such as technological

savviness and prior exposure to VR in shaping consumer engagement. Another significant gap lies in the impact of exposure sequencing whether transitioning from a simple VR environment to a detailed one or vice versa on user perceptions and behavior. Although traditional business studies have examined sequence effects in static advertisements, their implications in immersive VR settings remain largely unexplored. By investigating these gaps, this part of the research seeks to provide a nuanced understanding of how VR design complexity and exposure order can optimize consumer engagement and influence purchase decisions.

6.2.1 Problem Statement

To address these identified research gaps and integrate VR into the scope of this dissertation, which focuses on a systematic analysis of innovative technologies for consumer engagement in e-commerce, the following research questions are addressed in this chapter:

- **RQ 3.1:** How do individual differences, such as technological savviness, brand familiarity, and frequency of online gaming and shopping, moderate the relationship between VR complexity and engagement?
- **RQ 3.2:** How does designed visual complexity and opportunities for engagement within VR affect perceived immersion, engagement, realism, sense of presence, distraction, effort, and purchase intent?
- **RQ 3.3:** How do perceptions of VR interactions, such as perceived immersion, engagement, realism, etc., influence likelihood to purchase?
- **RQ 3.4:** What are the cognitive and emotional effects of exposure sequencing from a simpler VR to a more detailed VR environment and vice-versa?

6.3 Methodology

A desktop virtual reality study was performed. The virtual environments provided participants with advertisements of various Tesla and Apple products. The study had approval for human subjects research from the CSU Institutional Review Board (IRB).

6.3.1 Participants

There were 61 people that participated in the study. However, six participants were removed due to incomplete responses (i.e., only completed one or two of the surveys, not all three). After this data cleaning, there were a total of 55 participants included in the analysis. Participants were recruited from the undergraduate population at San José State University and received extra credit in one of their classes for participating in the study. Table 6.1 illustrates the distribution of participants across different age groups and genders involved in the study. The sample comprised of participants ranging from 18 to 50 years old.

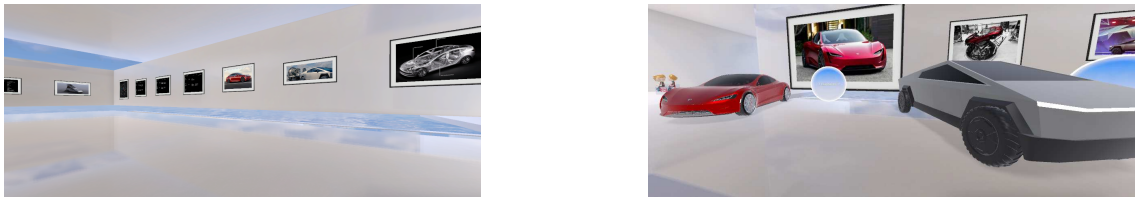
Table 6.1: Frequency Distribution of Gender Across Age Groups

Age Group	Male	Female
18-25	29	16
26-35	3	4
36-45	1	1
46-50	1	0

6.3.2 VR Environments

The experiment included two distinct virtual reality environments, both containing advertisements for the same Apple and Tesla products, representing a virtual Tesla showroom and Apple store. One VR environment, referred to as VR Simple (Figure 6.1a), provided participants with a basic immersive experience, where participants could walk around and watch 2-dimensional video advertisements about the products; designed to be minimalistic and informative. The other VR environment, referred to as VR Detailed (Figure 6.1b), provided the same products and videos, but

additionally displayed each product as a 3-dimensional model that enabled a more detailed visual experience; designed to be more visually stimulating. While we use the terms VR Simple and VR Detailed for use in this paper, the environments were referred to as VR Room A and VR Room B to the participants, to not bias them in their interactions.



(a) VR Simple **(b)** VR Detailed

Figure 6.1: Example of the two VR environments for the Tesla products.

A more detailed example of the VR Detailed environment with the Apple and Tesla products used in this study is provided in Figure 6.2. The featured Apple products include Apple Watch, Vision Pro, AirPods, iPhone, iPad, and Mac. Featured Tesla products include vehicle models S, 3, X, Y, and Optimus the Tesla Bot.

(a) Virtual Apple Store in VR Detailed. **(b)** Virtual Tesla Showroom in VR Detailed.

Figure 6.2: Example products in VR Detailed for the two brands.

These VR environments were developed and deployed using the Spatial platform. Participants were asked to create a free Spatial account in order to access the environments. We used the web-based version of Spatial, meaning interaction with the environments was done using a computer with a mouse and keyboard.

6.3.3 Participant Groups

Participants were assigned to one of two groups based on the first letter of their first name. This group determined the order in which they experienced the VR environments. Group Detailed-Simple (N = 33), assigned to those with first names beginning with letters A through O, interacted first with VR Detailed and subsequently VR Simple. Conversely, Group Simple-Detailed (N = 22), which included participants whose first names started with letters P through Z, experienced VR Simple first, followed by VR Detailed. This grouping strategy aimed to ensure a random distribution of participants and to control for potential confounding variables [214].

6.3.4 Data Collection Procedure

Data was collected in September and October of 2024. Participants were given written instructions on how to complete the study procedures and told to perform the study tasks on their own computers. The instructions provided links to the different VR environments and surveys. The participants' first and last names were collected on each survey so that responses could be linked across the different surveys. Each participant first completed a pre-exposure survey to capture baseline characteristics about them. They were then instructed to explore the first VR environment and spend approximately 5-minutes in the environment. Immediately after, they completed a post-exposure survey on that VR environment. Then, they were asked to explore the second VR environment and spend a similar amount of time in it. Lastly, they completed a post-exposure survey about the VR environment they just interacted with, which comprised of the same questions as the other post-exposure survey. Each survey captured various facets of the participants' experiences, including familiarity with products, interest in products, likelihood to purchase, and perceived engagement, immersion, and cognitive load of the VR environments.

Pre-Exposure Survey

The pre-exposure survey collected information on participant demographics and prior familiarity with the technologies and brands featured in the VR environments. Using Likert scale questions, we captured their self-reported tech-savviness, frequency of playing computer games, fre-

quency of shopping online, familiarity with VR, familiarity with Apple products, familiarity with Tesla products, interest in Apple products, and interest in Tesla products.

Post-Exposure Surveys

The post-exposure surveys were designed to capture participants' immediate reactions following their interactions with each virtual environment. The surveys were identical and were completed immediately after interacting with VR Simple and VR Detailed. Using Likert scale questions, participants were asked to rate how engaging, immersive, real, present, and distracting they found the experience. Questions also captured data on the physical and cognitive effort required to navigate the environments. Additionally, participants evaluated their interest in the products and their likelihood of purchasing the products, providing valuable information on how such an environment might influence consumer behavior and purchase intent.

6.3.5 Data Quality and Integrity

Data collection was conducted using structured online surveys administered at appropriate stages of the experiment. Responses were recorded using Likert scales across all datasets. Standard data cleaning procedures were implemented to handle missing values, verify consistency in response scales, and maintain participant anonymity and data integrity. In the survey, some Likert scale questions were reverse-coded to ensure participants were paying attention and ensuring data quality. These reverse-coded questions were then flipped for data analysis to ensure consistency in reporting across scales. Data cleaning and analysis were performed using both Python and R.

6.4 Results

6.4.1 Summary of Pre-Exposure Variables

A correlation analysis was first conducted on the pre-exposure variables related to prior experience with VR, gaming, online shopping, Tesla, and Apple. Figure 6.3 presents this correlation

heatmap depicting the pairwise relationships with Pearson correlation coefficients and statistical significance.

Figure 6.3: Correlation heatmap of pre-exposure variables.

The analysis reveals several significant correlations that provide insights into participants' technological behaviors and brand associations prior to the VR interventions. Two particularly strong positive correlations are observed between tech-savviness with VR familiarity ($r = 0.43, p < .01$) and with frequency of playing computer games ($r = 0.46, p < .001$). This indicates that participants who consider themselves more technologically savvy tend to have greater familiarity with virtual reality technology and more often play computer games. Additionally, there is a positive correlation between VR familiarity and gaming frequency ($r = 0.34, p < .05$), which suggests that participants who play computer games more frequently are also more familiar with VR technology. A noteworthy negative correlation is found between gaming frequency and Apple interest ($r = -0.33, p < .05$). This suggests that participants who frequently engage in computer gaming tend to have a lower interest in Apple products.

Other correlations of interest include the relation between Apple familiarity and Apple interest ($r = 0.62, p < .001$), which indicates that familiarity with Apple products enhances interest in the brand, which may influence participants' receptiveness to Apple products within the VR settings. A similar relation is seen between Tesla familiarity and Tesla interest ($r = 0.59, p < .001$). As well as interest in Apple and Tesla ($r = 0.43, p < .001$), although not significant for familiarity with Apple and Tesla ($r = 0.24, p > .05$).

Pre-Exposure Variables by Gender

Further analysis of these pre-exposure variables based on gender is presented in Figure 6.4. These pair plots illustrate the relationships among pre-exposure variables, where blue represents males and red represents females. The diagonal plots display the distribution of each variable by gender, highlighting differences and similarities in central tendencies and variability within each gender. Off-diagonal plots feature scatter plots with overlaid linear regression lines and kernel density estimates, which facilitates the exploration of potential linear relationships and data point distributions between pairs of variables. The Pearson correlation coefficient (ρ) is annotated on each scatter plot, denoting the strength and direction of the linear relationship between the variables.

A notable trend observed in Figure 6.4 is that gender differences only appear in the frequency of gaming variables. Wilcoxon rank-sum tests on each pre-exposure variable to compare responses between males and females verifies this; where males report gaming significantly more frequently than females ($p < .001$), while no other variable has significant differences between males and females ($p > .05$).

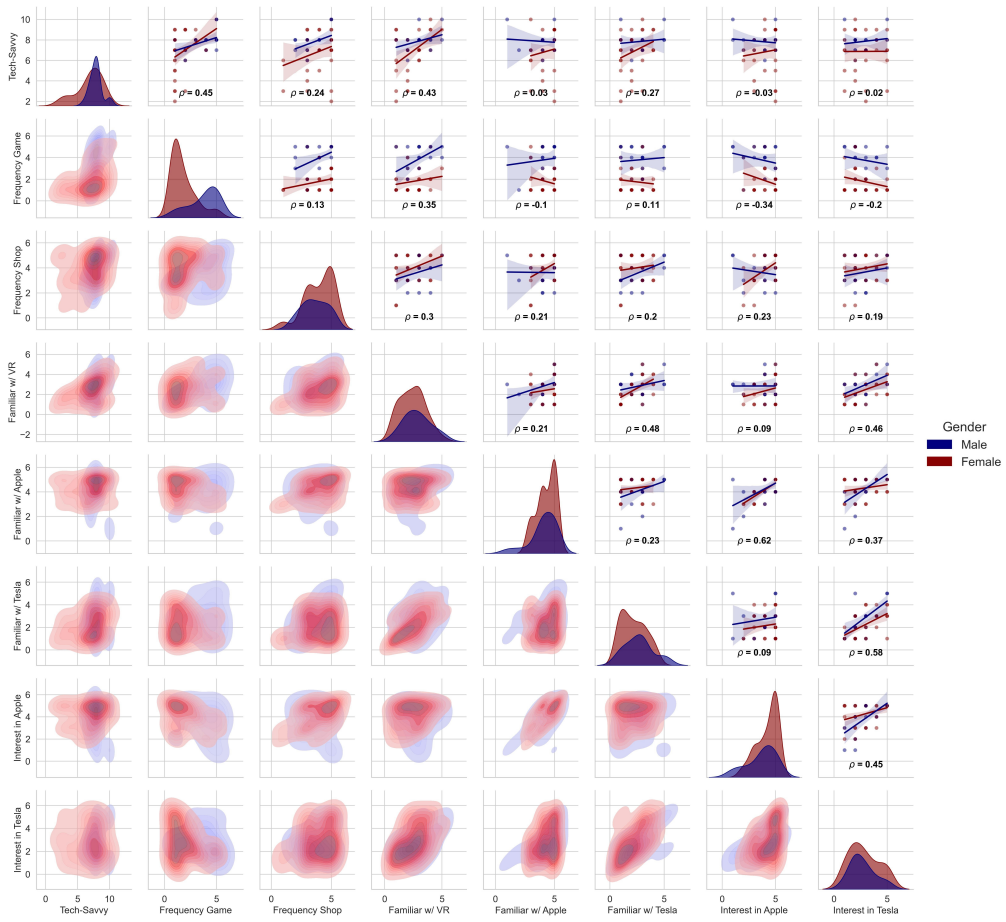


Figure 6.4: Pair plot of pre-exposure variables by gender.

6.4.2 Effects of Individual Differences on Perceived Engagement

An Ordinary Least Squares (OLS) linear mixed model on perceived engagement based on the pre-exposure variables is presented in Table 6.2. This model examines the influence of individual differences such as tech-savviness, brand familiarity, and behavioral frequencies on perceived engagement, controlling for the two VR environments. The model also includes a random effect for participants, to account for the repeated measures at the participant level (i.e., an observation for VR Simple and VR Detailed for each participant).

Most notably, the VR environment has a significant effect ($p = .001$) on perceived engagement, with VR Simple associated with an average engagement score of 0.854 less than VR Detailed. Additionally, brand familiarity similarly influences perceived engagement; where each unit increase

in familiarity with Apple increases engagement by 0.536 points ($p = .039$), and each unit increase in familiarity with Tesla increases engagement by 0.375 points ($p = .058$). Self-reported tech-savviness ($p = .555$), frequency of playing computer games ($p = .364$), and frequency of shopping online ($p = .109$) do not significantly impact perceived engagement.

Table 6.2: Summary of Linear Mixed Model Predicting Perceived Engagement

Variable	**Coef.**	**Std. Err.**	**t-value**	**p-value**
(Intercept)	2.625	1.396	1.880	0.065
VR Simple *(ref. VR Detailed)*	-0.854	0.245	-3.482	0.001
Tech-Savvy	-0.085	0.144	-0.593	0.555
Familiar w/ Apple Products	0.536	0.252	2.126	0.039
Familiar w/ Tesla Products	0.375	0.194	1.936	0.058
Frequency Computer Games	0.138	0.151	0.915	0.364
Frequency Shop Online	-0.329	0.201	-1.632	0.109
Model Fit	*AIC*	*LL*	*LRatio*	*p-value*
Model	447.97	-214.98	–	–
Null	449.08	-221.54	13.11	.041

Marginal Effects of Predictors on Perceived Engagement

Further analysis using partial dependence plots (PDPs) provides valuable insights into the relationships between perceived engagement and these individual predictors for both VR Simple and VR Detailed environments. While regression analysis provides information about the statistical significance of predictors, these PDP visualizations, which are often used in machine learning (i.e., Random Forest), interpret and offer a complementary perspective by exploring non-linear trends and marginal effects of each predictor while holding all other variables constant [215]. Specifically, these PDPs show how certain predictors influence perceived engagement under specific conditions, which might not be fully captured by linear models.

The theory behind PDPs can be described as, for a trained machine learning model $\hat{f}(\mathbf{X})$ that predicts an outcome y based on a set of features $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, the partial dependence function for a specific feature X_s is mathematically defined as:

$$\hat{f}_{X_s}(x_s) = \mathbb{E}_{\mathbf{X}_{-s}}\left[\hat{f}(x_s, \mathbf{X}_{-s})\right] \tag{6.1}$$

In this equation, x_s represents a fixed value of the feature X_s, while $\mathbf{X}_{-s}$ denotes the set of all other features except X_s. The operator $\mathbb{E}_{\mathbf{X}_{-s}}$ computes the expectation over the marginal distribution of $\mathbf{X}_{-s}$ and the partial dependence function $\hat{f}_{X_s}(x_s)$ represents the average model prediction when X_s is fixed at x_s, and all other features vary according to their marginal distributions [216].

We applied this to our linear model predicting perceived engagement, as shown in Figure 6.5 for VR Detailed and Figure 6.6 for VR Simple. As shown in Figure 6.5 for VR Detailed, the PDP for tech-savviness reveals a negative relationship with engagement, which indicates that individuals with higher technological expertise tend to perceive lower levels of engagement. This trend may reflect heightened expectations among more tech-savvy users, who are more open to innovative technologies but might show lower tolerance for simplicity or perceived shortcomings in VR experiences. Thus, it shows designers should consider more customizable features and interactions with more complex tasks. Conversely, familiarity with VR exhibits a positive trend, which suggests that individuals with prior exposure to VR are more likely to find Detailed VR environments engaging. This effect may stem from reduced cognitive load or greater ease of interaction for those already accustomed to VR technology. Also, familiarity with VR enhances the user's ability to focus on content rather than the interface. However, as VR becomes more common this effect may reduce, which highlights the dynamic challenges to sustain engagement among experienced users. Brand familiarity with Apple similarly shows a positive association with engagement while familiarity with Tesla presents a more nonlinear pattern where moderate familiarity is associated with higher engagement, but this effect diminishes at both low and high familiarity levels. This could be attributed to users with moderate familiarity likely balancing curiosity and comfort, while high familiarity may result in a loss of novelty, and low familiarity may lead to disinterest. The frequency of gaming appears to have minimal influence on engagement in VR Detailed, as the relationship remains relatively flat.

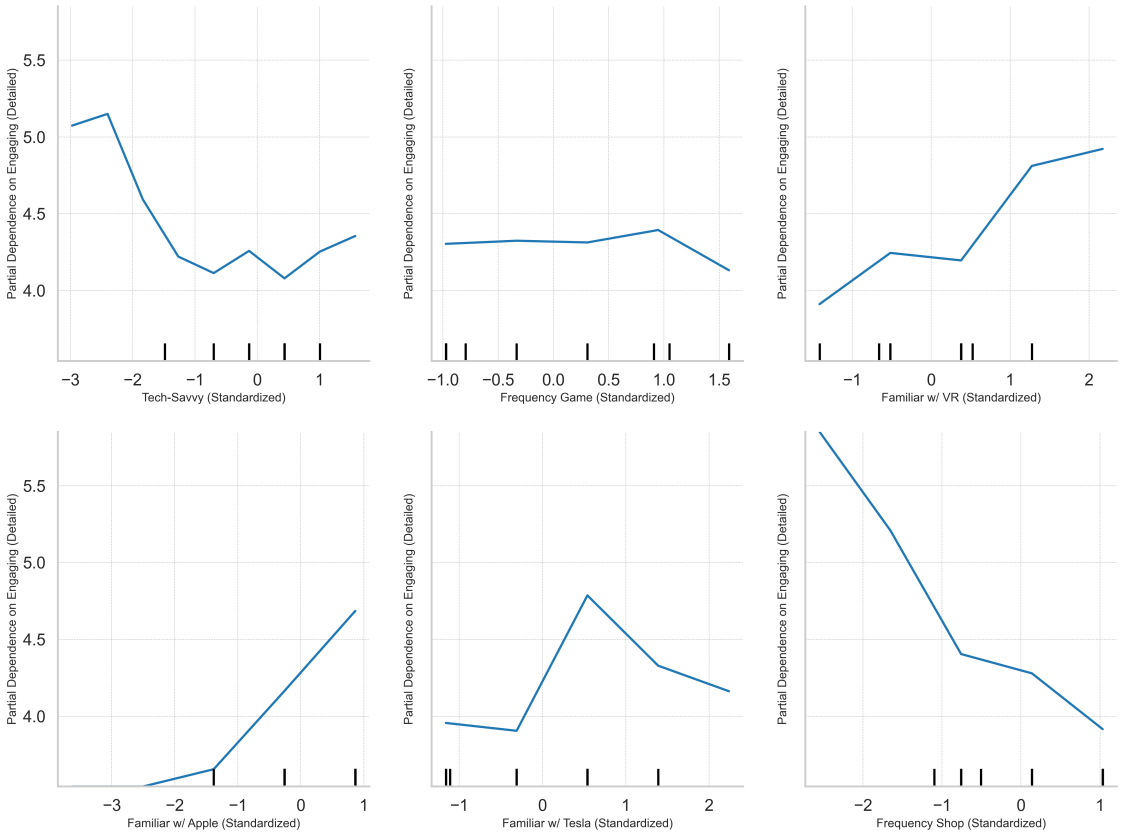


Figure 6.5: Partial dependence plots (PDPs) for predictors influencing perceived engagement in VR Detailed.

The VR Simple PDPs (Figure 6.6) reveal patterns that are both consistent with and distinct from those observed in the VR Detailed environment. The negative association between technological savviness and engagement persists, which reinforces the idea that higher expectations among tech-savvy individuals may reduce their engagement with simpler VR experiences. Familiarity with VR in VR Simple shows a fluctuating trend with engagement peaking at moderate levels of familiarity before declining among those with very high familiarity, potentially reflecting a novelty effect for users less experienced with VR technology. This suggests the need for adaptive VR experiences that evolve based on user expertise. Features like dynamic tutorials or content progression could help sustain engagement. Familiarity with Apple demonstrates a pronounced positive relationship and it highlights the role of brand affinity in shaping engagement in simpler virtual environments. Familiarity with Tesla exhibits a pattern similar to that in Detailed VR, where moderate familiarity correlates with the highest levels of engagement.

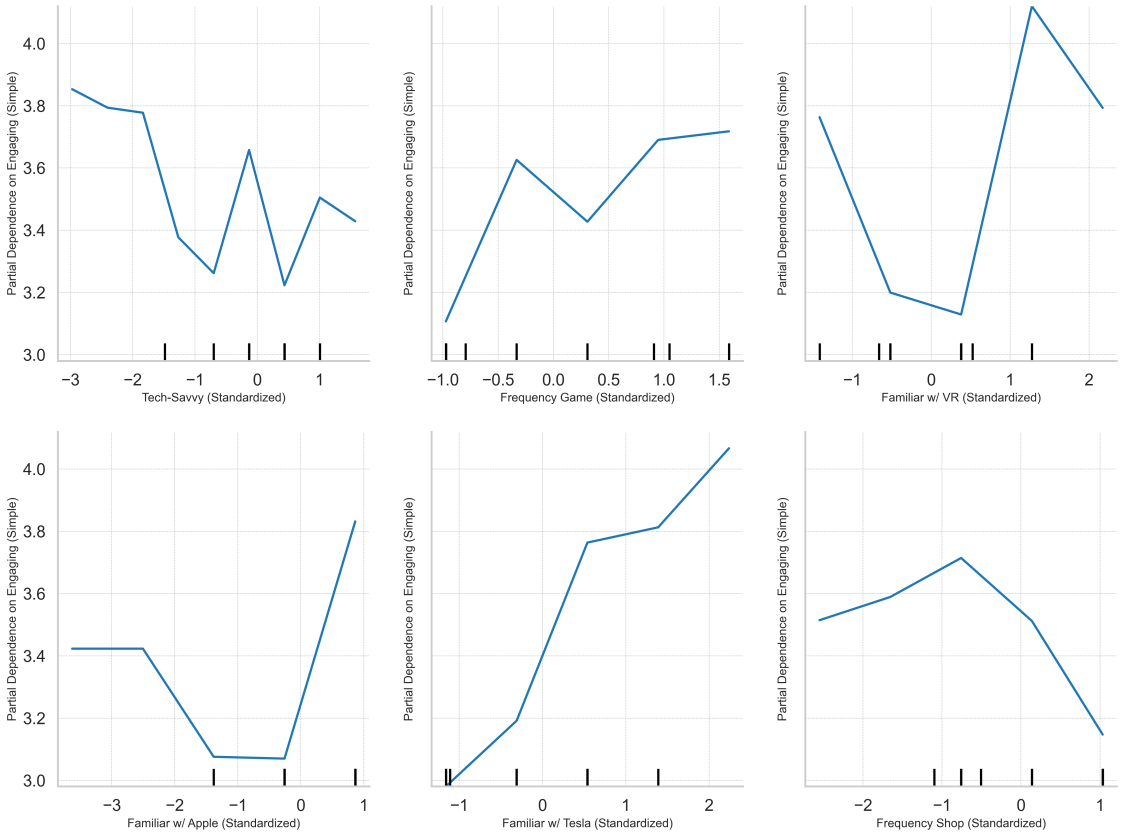


Figure 6.6: Partial dependence plots (PDPs) for predictors influencing perceived engagement in VR Simple.

6.4.3 Effects of VR Environment Design on Perceptions

A comparative analysis of post-exposure perceptions between VR Simple and VR Detailed, including perceived distraction, engagement, immersion, realism, effort, likelihood to purchase, and sense of presence, is provided in Figure 6.7. This comprehensive comparison sheds light on how the differing levels of designed complexity in the two virtual environments influence participants' experiences and perceptions. The density plots illustrate the distributions of participants' responses to the respective Likert scale questions for each VR environment. A paired Wilcoxon rank-sum test was performed for each scale between VR Simple vs VR Detailed, where these p-values are also provided in Figure 6.7.

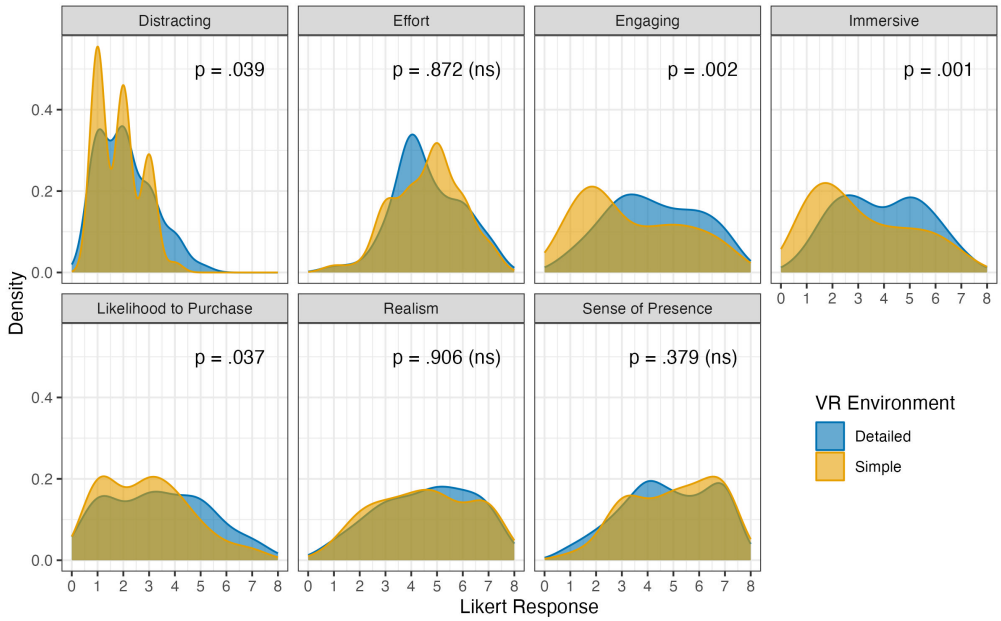


Figure 6.7: Density plots comparing post-exposure perceptions between VR Simple and VR Detailed.

The analysis reveals that VR Detailed generally outperforms VR Simple in terms of engagement ($p = .002$), immersion ($p = .001$), and likelihood to purchase ($p = .037$). Participants report higher levels of engagement in VR Detailed, which indicates that the dynamic and complex environment was more effective in capturing their attention and interest. The enhanced 3D model input and interactive elements in VR Detailed likely contribute to a more captivating experience, leading to increased participant involvement. Immersion scores are also higher for VR Detailed. This may stem from the detailed graphics and realistic 3D models and interactive features that VR Detailed offers, which can create a more convincing virtual experience and facilitate deeper cognitive and emotional engagement. The likelihood to purchase is also significantly higher following interaction with VR Detailed. This finding implies that the immersive features of VR Detailed not only enhance the immediate user experience but also positively influence participants' intentions to purchase the products featured. The increased likelihood of purchasing indicates that immersive VR environments can be powerful tools for influencing consumer behavior and driving sales.

It is also noteworthy that participants describe VR Detailed as significantly more distracting than VR Simple ($p = .039$). This suggests that distraction may not be viewed as a negative

attribute in such a setting, as participants felt more engaged, immersed, and likely to purchase in this more distracting environment.

However, metrics such as realism, effort, and sense of presence show similar ratings across both VR environments ($p > .05$). Participants perceive both VR environments as comparably realistic and even the less complex VR Simple effectively creates a believable virtual setting. The lack of significant differences in presence suggests that both environments are successful in making participants feel present within the virtual space. The effort required to interact with the VR environments does not differ markedly between VR Simple and VR Detailed. Participants do not perceive the more complex VR Detailed as requiring more effort, which is an important consideration for user experience design. An increase in complexity or immersion that does not add to the user's cognitive or physical burden is desirable, as it enhances engagement without causing fatigue or frustration.

6.4.4 Effects of VR Perceptions on Likelihood to Purchase

A linear mixed model was also fit to predict the likelihood to purchase based on participants' self-reported perceptions of immersion, engagement, realism, sense of presence, distraction, and effort. These results are presented in Table 6.3. The results show that the reported likelihood to purchase increases, on average, by 0.336 points for each unit increase in perceived engagement ($p = .009$). Similarly, each unit increase in realism increases the likelihood to purchase by 0.153 points ($p = .035$), while each unit increase in sense of presence decreases the likelihood to purchase by 0.2 ($p = .041$). Meanwhile, there is no effect of perceived effort, immersiveness, or distraction on the likelihood to purchase ($p > .05$).

Table 6.3: Summary of Linear Mixed Model Predicting Likelihood to Purchase

Variable	**Coef.**	**Std. Err.**	**t-value**	**p-value**
(Intercept)	0.846	0.894	0.947	0.348
Engaging	0.336	0.124	2.706	0.009
Realism	0.153	0.071	2.172	0.035
Sense of Presence	-0.200	0.095	-2.096	0.041
Effort	0.184	0.102	1.799	0.078
Immersive	0.169	0.141	1.199	0.236
Distracting	-0.075	0.139	-0.541	0.591
Model Fit	*AIC*	*LL*	*LRatio*	*p-value*
Model	383.13	-182.56	–	–
Null	422.53	-208.27	51.40	$< .0001$

6.4.5 Effects of Ordering on Perceived Engagement

To evaluate the causal effect of the VR exposure sequence/ordering on participants' engagement levels, we employed propensity score matching (PSM). PSM is a statistical technique that reduces selection bias in observational studies by balancing groups based on covariates predicting the likelihood of treatment. In this context, "treatment" refers to the sequence of VR room exposure, specifically, whether participants experienced VR Simple first, Group Simple-Detailed, ($T = 1$) or VR Detailed first, Group Detailed-Simple, ($T = 0$). The model evaluates perceived engagement in VR Detailed, hence whether exposure to the simpler environment first versus no prior environment to compare VR Detailed to, influenced perceived engagement in VR Detailed.

The propensity score is defined as the conditional probability of assignment to the treatment given a set of observed covariates, which in this case, the propensity score represents the probability that participant i receives the treatment (i.e., exposure to Simple first) given their covariates X_i:

$$e(X_i) = P(T_i = 1 \mid X_i) \tag{6.2}$$

where T is the treatment indicator, and X_i is the vector of observed covariates. In our model, the covariates include measures of tech-savviness, gaming frequency, online shopping frequency, VR familiarity, familiarity with Apple and Tesla products, age, and gender.

To estimate the propensity scores, we used a logistic regression model, represented as follows:

$$\text{logit}(e(X_i)) = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} \tag{6.3}$$

where $\text{logit}(e) = \ln\left(\frac{e}{1-e}\right)$ and β_0 is the intercept, β_k are the coefficients for each covariate X_{ik}, and p represents the total number of covariates. The logistic regression model was used only to calculate the propensity scores for matching purposes. It represents the probability of each participant being in the treatment group (Simple-Detailed) based on covariates such as measures of tech-savviness, gaming frequency, online shopping frequency, VR familiarity, and familiarity with Apple and Tesla products.

Following propensity score estimation, participants in the treatment group were matched to those in the control group using nearest-neighbor matching without replacement to ensure balance between the groups on observed covariates. For each participant i in the treatment group, we identified a corresponding participant j in the control group such that:

$$j = \arg\min_{j \in C} |e(X_i) - e(X_j)| \tag{6.4}$$

where C is the set of control participants.

This approach minimizes the absolute difference in propensity scores between matched pairs to ensure that the covariate distribution is as similar as possible between the two groups.

This technique specifically explores the causal effect of exposure sequencing on user engagement in VR environments to compare a Simple-Detailed sequence to a Detailed-Simple sequence. Thus, using PSM to reduce confounding variables, we found how exposure sequencing and design complexity interact to shape user engagement. As shown in Table 6.4, before matching, covariates exhibited varying degrees of imbalance, as indicated by an average Absolute Standardized Mean

Difference (ASMD) of 0.302. Post-matching, the mean ASMD dropped to 0.144, with 85.7% of covariates achieving ASMD < 0.2 and $28.6\% < 0.1$, which reflects substantial improvements in balance. Before matching, treatment and control groups exhibited significant divergence in propensity scores. Post-matching, the overlap improved substantially, validating the matching approach. The ASMD formula is shown below, where μ_t and μ_c are the means for the treatment and control groups and σ_t^2 and σ_c^2 are the variances for the treatment and control groups, respectively.

$$\text{ASMD} = \frac{|\mu_t - \mu_c|}{\sqrt{\frac{\sigma_t^2 + \sigma_c^2}{2}}} \tag{6.5}$$

Table 6.4: Summary of ASMD Values Before and After Matching, and Reduction Percentages for Each Covariate

Covariate	**Before**	**After**	**Reduction (%)**
Familiar w/ Apple Products	0.125	0.129	-2.65
Familiar w/ Tesla Products	0.046	0.000	100.00
Familiar w/ VR Technology	0.182	0.044	75.99
Frequency Computer Games	0.683	0.219	67.86
Frequency Shop Online	0.134	0.137	-2.38
Gender (Numeric)	0.219	0.109	50.12
Tech-Savvy	0.432	0.188	56.45

After performing the matching, we obtained a sample in which the distribution of the measured covariates was more closely aligned between the treatment (Simple-Detailed) and control (Detailed-Simple) groups. This allowed us to more confidently attribute differences in engagement to the sequence of exposure rather than underlying participant characteristics.

Using the matched sample we analyzed the key outcomes of interest, specifically engagement scores in VR Detailed (Y_B) for each VR environment. We estimated the average treatment effect on the treated (ATT), formulated as:

$$\text{ATT} = \frac{1}{N_T} \sum_{i \in T} \big(Y_i(1) - Y_{m(i)}(0)\big) \tag{6.6}$$

where N_T represents the number of treated participants, $Y_i(1)$ is the outcome for treated participant i, $Y_{m(i)}(0)$ denotes the outcome for the matched control participant $m(i)$, and T indicates the set of treated participants. The ATT thus measures how much higher (or lower) engagement scores are for those who saw the Simple VR environment first, compared to what their engagement on average would have been if they had not.

To investigate the impact of VR exposure sequence on engagement levels, we generated a kernel density estimate (KDE) plot shown in Figure 6.8. This plot compares the distribution of engagement scores for VR Detailed based on participants in Group Simple-Detailed versus Group Detailed-Simple. This plot shows the actual engagement scores for the VR Detailed after matching participants using propensity scores. In the plot, the x-axis represents the self-reported engagement scores for VR Detailed, as measured on a standardized Likert scale, while the y-axis denotes the probability density function estimated using kernel density estimation. The KDE smoothing process can result in a range of values extending beyond the original scale of 1 to 7 as this is an artifact of the smoothing technique and does not imply the engagement score outside the actual range. The density plot reveals a noticeable rightward shift in the engagement score distribution for the treatment group (VR Simple-Detailed) compared to the control group (VR Detailed-Simple). Specifically, the peak for Simple-Detailed occurs at a higher engagement score compared to the peak engagement score for the Detailed-Simple group. Thus, this plot indicates that participants who experienced Simple VR followed by Detailed VR reported higher engagement levels in Detailed VR than those who experienced Detailed VR without prior exposure to Simple VR. Moreover, the treatment (Simple-Detailed) group's density curve exhibits a higher concentration of participants with engagement scores near the upper end of the scale.

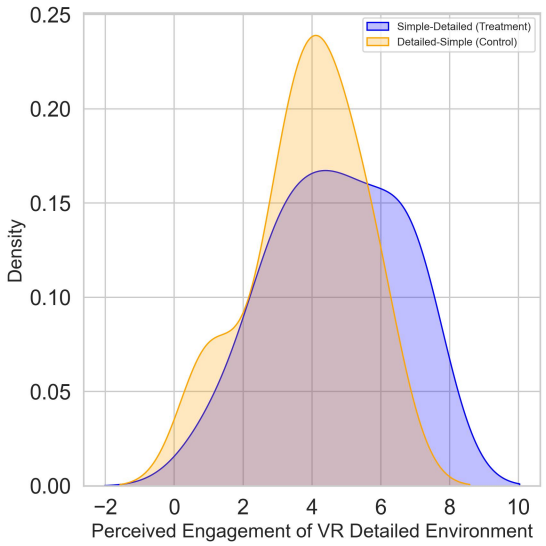


Figure 6.8: Density plot of VR Detailed engagement scores by treatment group for propensity matching scores.

To rigorously assess the statistical significance of the observed differences in engagement scores in VR Detailed between the treatment and control groups, we conducted a two-sample Kolmogorov-Smirnov (K-S) test [217]. The K-S test evaluates the maximum absolute difference between the empirical cumulative distribution functions (ECDFs) of the two samples. The test statistic D is defined as:

$$D = sup_x \left| F_{\text{treatment}}(x) - F_{\text{control}}(x) \right| \tag{6.7}$$

where $F_{\text{treatment}}(x)$ and $F_{\text{control}}(x)$ are the ECDFs of the engagement scores for the treatment and control groups and $\sup_x$ denotes the supremum over all possible values of x.

In our analysis, the K-S test yielded a test statistic of $D = 0.456$ with a corresponding $p = 0.008$. Since the p-value is less than the conventional significance level of $\alpha = 0.05$, we reject the null hypothesis that the engagement scores in VR Detailed for the treatment and control groups come from the same distribution. This result suggests a statistically significant difference in the engagement distributions attributable to the exposure sequence.

Also, to illustrate these differences, we present a cumulative distribution function (CDF) plot in Figure 6.9. This diagram visualizes the empirical CDFs of the engagement scores for both the treatment and control groups. The horizontal axis represents the range of engagement scores in VR Detailed, while the vertical axis denotes the cumulative probability. As shown in Figure 6.9, the empirical CDF of the treatment group (Simple-Detailed) lies consistently to the right of the control group's CDF (Detailed-Simple) and it indicates higher engagement scores among participants who experienced Simple VR before Detailed VR. The maximum vertical distance between the two curves corresponds to the K-S statistic $D = 0.456$ (the largest discrepancy between the two distributions). The significant difference in the distributions of engagement scores, as evidenced by both the K-S test and the CDF diagram, highlights the influence of exposure sequence on user engagement.

These analyses are derived from a predictive model rather than raw observational data. By incorporating covariates and controlling for cofounders through both the logistic regression for propensity score calculation and the subsequent matching, the differences we observe are more likely to represent the true effect of the VR exposure sequence rather than artifacts of pre-existing differences between participant groups.

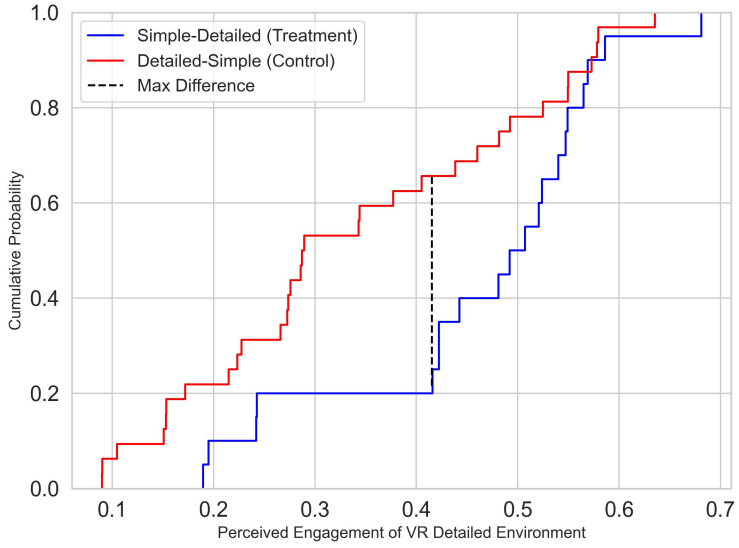


Figure 6.9: Empirical cumulative distribution functions of VR Detailed engagement scores by treatment group.

6.5 Discussion

The results of this chapter provide new insights into the impact of virtual reality (VR) design complexity and exposure sequencing on consumer engagement and purchase intent, while accounting for individual characteristics such as technological savviness and brand familiarity. Our findings align with and extend existing research in immersive marketing and user experience design.

The demographic composition of the participants is a crucial factor in interpreting the study's findings. Since the sample is predominantly young and female, the results may reflect the perceptions and behaviors typical of this demographic segment. Younger individuals are often more receptive to emerging technologies like virtual reality and may exhibit different engagement levels compared to older cohorts [218, 219]. Similarly, gender-related preferences and familiarity with brands such as Apple and Tesla could impact participants' interactions within the VR environments. Similarly, those familiar with gaming have likely skilled their spatial and interactive skills through virtual worlds, and they may more readily adapt to VR scenarios [220].

Analysis of the pre-exposure variables highlights that while our participant group is uniformly tech-savvy, their engagement in specific technology-related activities such as gaming and online shopping varies considerably. The high level of tech-savviness across the sample is advantageous for the study, as it suggests that participants are likely comfortable navigating virtual reality environments. However, the variability in gaming frequency may influence individual familiarity with interactive digital environments, potentially affecting immersion and engagement levels within the VR settings. Similarly, differences in online shopping frequency could impact participants' responsiveness to product exploration tasks and purchase intent within the virtual environments.

The majority of our participants reported fairly high familiarity with Apple's products, with limited variability. This high familiarity with Apple products among participants is consistent with Apple's widespread market presence and the prevalence of its devices in everyday life [221]. Familiarity with Tesla products showed that while many participants are familiar with Tesla products, there is greater variability in their familiarity levels. The distribution suggests that some partici-

pants are highly familiar with Tesla, while others have limited exposure. The broader range may reflect the more specialized nature of Tesla's product line, which includes Optimus (the robotic humanoid) and energy solutions that may not be as widely encountered as consumer electronics from Apple. The disparity in interest levels between Apple and Tesla may be attributed to several factors. Apple, as a consumer electronics giant, has a pervasive presence in the daily lives of many individuals through its extensive product lineup, including iPhones, iPads, MacBooks, and related services. This ubiquity likely contributes to the higher and more consistent interest levels observed among participants, many of whom may be regular users of Apple devices or services. Tesla, on the other hand, specializes in electric vehicles and sustainable energy products, which, while innovative, may not be as accessible or relevant to the participants, particularly given the demographic composition of the sample. The participants, predominantly university students aged 18-25, may have limited financial means or immediate need for such high-investment products, which could explain the lower median interest and greater variability.

The lower familiarity with VR technology could have implications for the study, as participants' prior exposure to VR may influence their comfort levels, engagement, and immersion within the VR environments used in the experiment. Participants with limited familiarity may require more time to acclimate to the VR setting, which could affect their overall experience and responses in the post-exposure surveys. Flavian (2019) [222] on VR usage suggests that as users gain experience they often report higher levels of comfort and presence which highlights the importance of considering prior experience as a moderating factor. Overall, the disparity in familiarity levels across VR, Apple, and Tesla highlights the importance of considering participants' prior experiences with these technologies and brands. The varying degrees of familiarity may serve as moderating factors affecting how participants perceive and interact with the VR environments, as well as their subsequent engagement, immersion, and consumer behavior outcomes measured in the study.

We observed a strong positive correlation between technology savviness with VR familiarity and with frequency of playing computer games. The strength of these correlations suggests that

general technological proficiency is closely linked to exposure to or interest in VR and gaming, which may influence participants' comfort levels and engagement within the VR environments used in the study. Similarly, participants in our study who reported greater VR familiarity also reported higher gaming frequency. This is plausible given that VR has been increasingly integrated into gaming experiences [223]. Interestingly, we observed a negative relation between gaming frequency and Apple interest. This relationship may stem from the perception among some gamers that Apple's products are less suited for high-performance gaming compared to other platforms since Apple's focus on design and user experience may not align with the preferences of frequent gamers who prioritize hardware specifications and compatibility with a wide range of gaming software [224]. This negative correlation highlights the diversity of technological preferences among participants and may influence how different subgroups respond to Apple-related content in the VR environments. Understanding these divergent preferences is critical for understanding VR marketing interventions. For instance, content that highlights Apple's ecosystem and everyday utility may appeal less to dedicated gamers and more to generalist tech-savvy consumers whereas Tesla's innovative offerings may resonate with tech enthusiasts interested in cutting-edge engineering and sustainability.

6.6 Conclusions

This chapter evaluated how the designed complexity of a desktop virtual reality environment and pre-exposure individual characteristics influenced consumer engagement, distraction, purchase intent, and product exploration behaviors. We conducted a 2 (VR design: Simple, Detailed; within-subject) x 2 (Order: Simple-Detailed, Detailed-Simple; between-subject) mixed factorial design study with 55 participants. Our VR environments contained videos of Tesla and Apple products (VR Simple), as well as 3D models of each product (VR Detailed).

Specifically, RQ 3.1 sought to understand how individual differences, such as technological savviness, brand familiarity, and frequency of online gaming and shopping, influence perceived engagement in a VR environment. Our results show that a more complex VR environment and greater

pre-familiarity with products resulted in increased perceived engagement, while tech-savviness and frequency of playing computer games and shopping online had no effect.

In RQ 3.2, we evaluated how designed visual complexity and opportunities for engagement within VR affect perceived immersion, engagement, realism, sense of presence, distraction, effort, and purchase intent. Our analysis demonstrated that the more complex VR environment was characterized by participants as more distracting, more engaging, more immersive, and led to a higher likelihood to purchase the products, as compared to the simpler VR environment. Whereas participants reported no difference in effort, realism, or sense of presence between the two VR environments.

RQ 3.3 explored how perceptions of the VR environment influenced the purchase intent of the Tesla and Apple products. Using a linear mixed model, our results identified that perceived engagement and realism increased the likelihood to purchase, sense of presence decreased the likelihood to purchase, effort navigating the VR environment marginally increased the likelihood to purchase, while immersiveness and distractions did not have an effect.

Lastly, our RQ 3.4 investigated how the sequencing of VR exposure affected perceived engagement using propensity score matching. We compared perceived engagement scores for the more detailed VR environment between those who had experienced a simpler VR environment before those who had not experienced any VR environment prior. The results support that perceived engagement is higher after being exposed to a simpler VR environment, rather than no prior exposure.

In conclusion, this research highlights the critical role of designed complexity, individual characteristics, and exposure sequencing in shaping consumer experiences within virtual reality environments. The findings demonstrate that more complex VR environments significantly enhance engagement and perceived realism without increasing cognitive load, providing valuable information to optimize user experience design. Additionally, the progression from simpler to more detailed VR settings amplifies engagement, suggesting a strategic approach to presenting VR content. By integrating these elements, businesses can effectively leverage VR technology to foster

consumer engagement, enhance brand perception, and influence purchase intent. These insights contribute to the growing understanding of VR as a transformative tool in consumer behavior and marketing strategies, paving the way for future research and innovation in immersive technologies.

Chapter 7

Aim 4: Extending Vision Deep Learning Models to VR

7.1 Aim 4 Summary

Virtual reality is changing e-commerce by providing immersive and more interactive shopping experiences. However, evaluating user engagement within these environments remains a challenge. This chapter uses the convolutional neural network (CNN) model validated in Research Aims 1 and 2, to analyze user engagement in the desktop VR study evaluated in Research Aim 3, where participants interacted with the virtual Apple store and virtual Tesla showroom. In this chapter, we propose different video-level interest probability using frame-level probability aggregation techniques such as mean, median, Gaussian-weighted, peak-weighted, and trend-aware attention methods to transition from frame-level to video-level engagement classification and overcome constraints in computational resources needed. The findings show that temporal aggregation significantly enhances predictive accuracy, with trend-aware fusion outperforming other methods by effectively capturing temporal dynamics. This research provides actionable information for optimizing VR commerce experiences and enabling businesses to incorporate interactions based on user engagement patterns. The results highlight the potential of AI-powered methodologies in refining VR user experience design and improving engagement-driven business strategies.

7.2 Introduction

Virtual reality is reshaping how users engage with digital environments, particularly in e-commerce, education, healthcare, industry, and entertainment. Unlike traditional online shopping which relies on static images and text, VR enables consumers to interact with products in a fully immersive 3D space, creating a shopping experience similar to visiting a physical store [225]. Ac-

curately measuring engagement in VR remains a challenge, but advances in deep learning, such as facial emotion recognition, provide a new way to evaluate engagement by analyzing user expressions in real-time [226].

Several studies have explored the use of FER in virtual environments. Caserman et al. [227] examined body tracking in VR to assess user motion, but their work focused on movement rather than facial expressions. Jeong et al. [228] introduced a multimodal model for predicting cybersickness in VR, considering factors such as physiological signals and head movement; while their approach offers valuable information into user experience, it is designed to prevent discomfort rather than analyze engagement.

Morin et al. [225] investigated self-avatars in mixed reality, demonstrating that personalized avatars enhance user presence and interaction. However, their study did not focus on emotion recognition. Meanwhile, Cortes et al. [226] conducted a review of CNN applications in VR to emphasize object tracking and interaction, but not to evaluate user engagement.

Despite these advancements, two major gaps remain in the body of knowledge regarding VR engagement. First, many studies rely on self-reported data or indirect behavioral metrics, which do not provide real-time insights into user emotions [229]. Second, FER in VR is prevented by occlusion issues caused by head-mounted displays, which obscure key facial features needed for accurate emotion detection [225].

This chapter expands upon prior work by integrating FER models using convolutional neural networks into VR commerce environments to analyze user interest in real-time. Unlike previous studies that focus on static engagement metrics [228], our approach uses frame-level probability aggregation to enhance classification accuracy over time.

Building on the Xception model from Research Aim 1 [4], our research incorporates temporal-based classification methods of video frames, rather than treating frames as independent observations, to extract meaningful engagement patterns from VR interactions. In this chapter, we compare the accuracy of the Xception CNN in classifying interested/not-interested using frame-level (baseline), mean, median, gaussian-weighted, peak-weighted, and trend-aware attention methods.

7.2.1 Problem Statement

Virtual reality is changing the landscape of e-commerce and is providing consumers with a new way to experience products and connect with brands [230]. VR environments create an immersive and interactive shopping experience that brings online retail closer to the feel of an in-store visit, which offers a richer and more engaging way for consumers to explore products [231]. However, the success of VR based e-commerce depends on how well it engages users and resonates with their emotions, as that plays a key role in shaping purchase decisions and the overall shopping experience [232]. While VR holds great promise, measuring user engagement in these environments remains a challenge. Traditional metrics like click-through rates [233] and dwell time [234] fall short of capturing the deeper emotional and cognitive interactions that make VR experiences unique. Recent advances in vision deep learning FER offer a promising way to handle this challenge. By using convolutional neural networks to analyze facial expressions, FER can classify emotional states and provide deeper information into user engagement [235]. Vision deep neural networks' potential in dynamic and interactive VR environments is largely unexplored. Unlike traditional settings, VR presents unique challenges such as real-time interactions which demand more advanced methods capable of accurately detecting user interest and disengagement as they occur.

To bridge this gap, Research Aim 4 focuses on extending vision deep learning models into immersive VR environments to analyze emotional responses and engagement patterns during virtual reality e-commerce interactions. This research expands upon the validated and tested Xception model from Research Aim 1 [4] by developing frame-level probability aggregation techniques to account for temporal effects, in an effort to improve the robustness of AI-based image detection in VR. This chapter seeks to address one core research question:

- **RQ 4.1:** How does accuracy compare for frame-level versus temporal-level CNN-based FER models in classifying user interest/disinterest in interactive VR environments?

7.3 Data Collection and Description

The data collected with the San José State University students, experimental design described in Research Aim 3 (Section 6.3), was also used to support this Research Aim 4. Recall, that in this study, a desktop virtual reality study was conducted to capture user engagement in two VR showrooms (Simple and Complex) for Apple and Tesla products.

7.3.1 Procedures

Participants were provided written instructions on how to complete the study procedures. Half of the participants were instructed to start with VR Simple and then do VR Complex, while the other half experienced the VR environments in reverse order. The VR rooms were described as VR Room A and VR Room B to the participants, so as to not bias them towards the respective designs. Each participant completed the study on their own time and with their own equipment. They were instructed to interact with each VR room on their computer and to record a video of their face for the entirety of their interaction with the VR room. Participants were asked to record this video using either their phone or their laptop, and then upload their video file. After each VR room, they completed a survey about the VR experience they just had. One of the questions in this survey was "Were you interested in the experience? (Yes/No)". Their self-reported responses from this survey question were used to tag their video as "interested" or "not-interested." Participants were asked to provide their first and last name in each survey and video file, to facilitate linking survey responses to videos.

Data collection occurred between September to October of 2024. Figure 7.1 provides an overview of the study procedures. Participants interacted with the desktop VR while providing video recordings of their faces, which were then used to compare CNN classification techniques for their ability to make engagement (interested/not-interested) predictions.

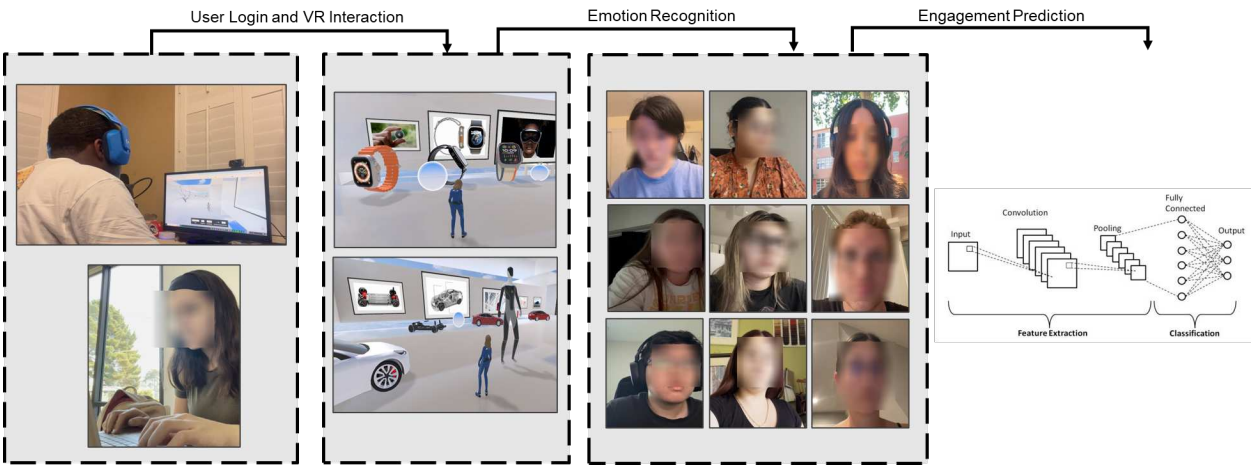


Figure 7.1: Experiment workflow for vision deep learning comparisons of user engagement in VR.

7.3.2 Participants

There were a total of 61 people who participated in the study. However, not every participant uploaded their video files. Hence, for this data analysis, there were 46 participants included. Table 7.1 summarizes the age and gender demographics for the participants, where there were a total of 19 males and 27 females, ranging in age from 18 to 45.

Table 7.1: Participant Ages and Genders

Age Group	Male	Female
18-25	14	23
26-35	4	3
36-45	1	1

7.3.3 Video Data

Initially, there were 95 different reaction videos from participants. However, after matching video filenames with survey responses, the final dataset included 91 videos, as 4 video files could not be matched to survey data. This included 45 videos from VR Simple and 46 videos from VR Complex.

The videos varied in duration (mean = 132.972 seconds, SD = 49.998 seconds) and frame rate (mean = 35.129 fps, range = 14.333 to 60.003 fps). To ensure computational efficiency and eliminate redundancy, a standardized frame sampling strategy was applied to extract one frame for every ten frames. After pre-processing, 42,712 frames were selected for FER analysis, which represents approximately 10% of the total frames.

Each frame was assigned a label of "Interested" or "Not Interested" based on the survey response. Key statistics of the videos and frames used for FER analysis are summarized in Table 7.2.

Table 7.2: Summary Statistics of the Dataset

Metric	Mean	Std Dev	Min	Max	N
Video Frame Rate (fps)	35.12	13.74	14.13	60.00	–
Video Duration (seconds)	132.97	49.99	12.30	398.14	–
Video Duration (frame count)	469.37	245.16	36.90	1,375.90	–
Videos Labeled "Interested"	–	–	–	–	59
Videos Labeled "Not Interested"	–	–	–	–	32
Frames Labeled "Interested"	–	–	–	–	29,584
Frames Labeled "Not Interested"	–	–	–	–	13,128

7.4 Analytical Methodology

The same CNN Xception architecture model that we trained, validated, and tested in Research Aim 1 [4] was used with this data. Recall, that this pre-trained model was fine-tuned for facial emotion recognition. Specifically, all 42,712 frames from this present dataset were passed through our Xception model of facial recognition to predict if the participant was interested or not interested.

While in Research Aim 1 we primarily used frame-level classification, this chapter expands upon the Xception methodology by incorporating frame-level probability aggregation to video-level (temporal) classification techniques that aggregate frame-level probabilities into a single and temporally consistent video-level probability. This refinement helps make the final classification better align with user engagement trends over time rather than relying solely on independent frame

predictions. Hence, we are able to account for the sequential video nature of the data, rather than treating each frame as independent observations. The classification methods used in this chapter are summarized in Table 7.3.

Table 7.3: Frame and Temporal Level Classification Methods Employed

Method	Type	Description
Frame (Baseline)	Frame	Each frame treated independent
Mean	Temporal	Average across frames in video sequence
Median	Temporal	Median across frames in video sequence
Gaussian-Weighted	Temporal	Higher weight to frames near center of video sequence
Peak-Weighted	Temporal	Higher weight to frames with strong expression intensity
Trend-Aware Attention	Temporal	Peak-weighting plus how expression changes over time

To transition from frame-level predictions to a single video-level interest probability, we employed multiple aggregation techniques, as described above. The formulas for each of these techniques are provided below.

Frame (Baseline) Method: The baseline method used in Research Aims 1 and 2, where each frame of a video is processed independently by the model, regardless of the temporal associations across frames from the same video.

Mean Method: A basic method of equal weight across all frames, that computes the average of classifications across all frames in a video sequence. See Equation 7.1.

$$P_{\text{video}} = \frac{1}{N} \sum_{i=1}^{N} p_i \tag{7.1}$$

where p_i is the CNNs classification for the i^{th} frame across the total number of frames (N), and P_{video} is the classification probability assigned to the video sequence.

Median Method: A basic method also applying an equal weight across all frames, that computes the median of feature classifications across all frames in a video sequence. See Equation 7.2.

$$P_{video} = \text{median}(p_1, p_2, \ldots, p_N) \tag{7.2}$$

where, $p_1 ... p_N$ are the CNNs classifications for the i^{th} frame, and P_{video} is the classification probability assigned to the video sequence.

Gaussian-Weighted Method: A somewhat more complex approach that applies a Gaussian distribution over time, giving higher weight to frames near the center of the video sequence. See Equation 7.3.

$$w_i = \exp\left(-\frac{(i-M)^2}{2\sigma^2}\right), \quad P_{video} = \frac{\sum_{i=1}^{N} w_i p_i}{\sum_{i=1}^{N} w_i} \tag{7.3}$$

where $M = N/2$ is the middle frame index, σ controls the spread of weighting for Gaussian weighted method, and the final video-level classification probability is P_{video}.

Peak-Weighted Method: A more complex approach that assigns higher weights to frames within a video sequence where the CNN's expression intensities are strongest, thus prioritizing the most expressive moments in a video. See Equations 7.4.

Formally, a frame i is considered a peak if:

$$w_i = f(p_i > p_{i-1} \text{ and } p_i > p_{i+1}), \quad P_{video} = \frac{\sum_{i=1}^{N} w_i p_i}{\sum_{i=1}^{N} w_i} \tag{7.4}$$

where f is an indicator function. This method prioritizes high-confidence moments by assigning weights based on local maxima in probability values.

Trend-Aware Attention Method: An adaptive approach, which is an extension of the peak-weighted approach that further accounts for temporal trends. Specifically, this approach assigns importance to frames based on the intensity of the expression and whether the expression is increasing, decreasing, or stable over time. See Equations 7.5, 7.6, 7.7, and 7.8.

The trend score is computed as:

$$T_i = |p_{i+1} - p_i| \tag{7.5}$$

The combined score S_i is calculated as:

$$S_i = \alpha p_i + (1 - \alpha) T_i \tag{7.6}$$

Frames are selected based on the highest combined scores:

$$\mathcal{F} = \{i \mid S_i \text{ in top } 50\% \text{ of all frames}\} \tag{7.7}$$

The final classification probability is given by:

$$P_{\text{video}} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} p_i \tag{7.8}$$

7.5 Results and Discussion

7.5.1 Receiver Operating Characteristic (ROC) Curves for Each Method

The six different classification methods (one frame-level, 5 video-level) were each used within the Xception CNN model for classifying the data. Figure 7.2 compares the performance, using Receiver Operating Characteristic curves, of these different methods for aggregating frame-level probabilities into video-level predictions. The ROC curve evaluates the trade-off between the true positive rate (sensitivity) - y-axis, and false positive rate - x-axis, with the area under the curve summarizing each method's performance. Higher AUC values indicate better discrimination between "Interested" and "Not Interested" classes. It is noteworthy that the trend-aware attention method achieved the highest AUC (0.927).

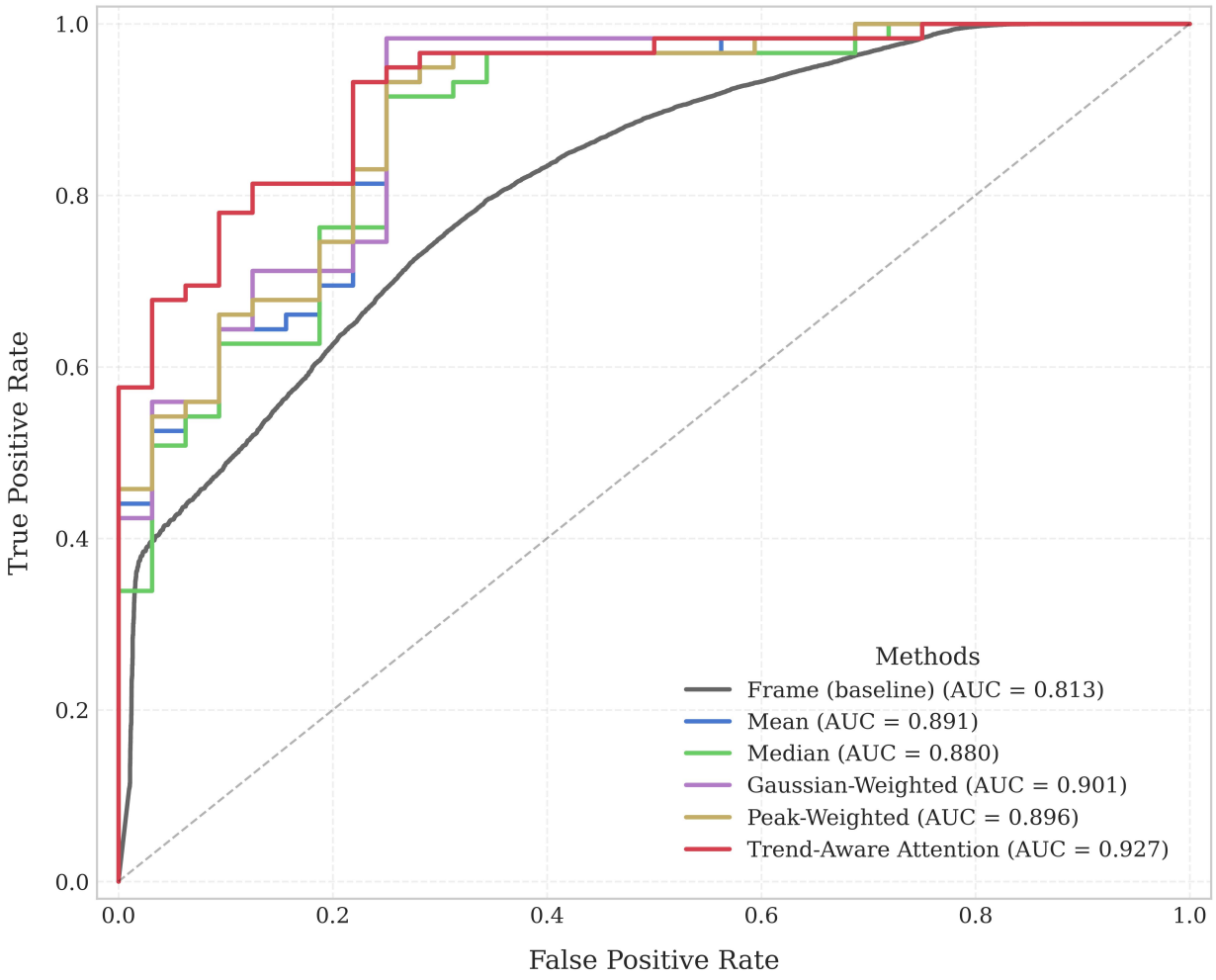


Figure 7.2: ROC curves comparing the six methods for aggregating frame-level probabilities.

Not surprisingly, the *frame*-level (*baseline*) predictions method, which treats each frame independently, yielded the lowest AUC (0.813). The simple *mean* approach, which averages frame probabilities, had improved AUC (0.891), but still does not fully account for temporal dependencies. Similarly, the *median* aggregation achieved a slightly lower AUC (0.880) compared to the *mean* approach, as the *median* approach focuses on the central tendency, but disregards dynamic frame relationships.

The more advanced methods of *Gaussian-weighted* (AUC = 0.901) and *peak-weighted* (AUC = 0.896) incorporate temporal weighting schemes, increasing performance compared to the *frame (baseline)* method. *Gaussian-weighted* emphasizes central frames, while *peak-weighted* prioritizes

moments of high confidence. However, these methods still fall short of capturing complex temporal trends fully.

The temporal fusion via *trend-aware attention* achieved the highest AUC (0.927), showing its superior performance. This method combines frame-level probabilities with temporal gradients, dynamically emphasizing frames with both high probabilities and significant temporal changes. By effectively capturing temporal dependencies and trends, this approach provides a more accurate representation of video-level interest.

7.5.2 Overall Performance for Each Method

In addition to the comparison of area under the curve (AUC) from the ROC curves, performance was compared across the methods based on their accuracy, precision, recall, and F1 score. Where accuracy describes the proportion of correct classifications across all *observations*; precision represents the correctly predicted positive classifications out of all *predicted positives*; recall describes the proportion of predicted positive classifications out of all *actual positives*; and F1 score is the harmonic mean of precision and recall.

This performance comparison across the different aggregation methods is provided in Table 7.4. As evidenced in this table, trend-aware attention achieved the highest performance in terms of accuracy (0.87), recall (0.95), F1 score (0.90), and AUC (0.93). Whereas the frame (baseline) and simple mean and median aggregations exhibit lower recall and overall performance; further showing their limitations in capturing temporal dependencies. The Gaussian-weighted and peak-weighted approaches perform better, but still worse than trend-aware attention, which dynamically integrates temporal patterns for improved classification accuracy.

Table 7.4: Overall Performance Comparison of the Different Aggregation Methods

Method	Accuracy	Precision	Recall	F1 Score	AUC
Frame (baseline)	0.706	0.85	0.68	0.756	0.813
Mean	0.74	0.87	0.69	0.77	0.89
Median	0.70	0.88	0.63	0.73	0.88
Gaussian-Weighted	0.76	0.89	0.71	0.79	0.90
Peak-Weighted	0.77	0.87	0.76	0.81	0.90
Trend-Aware Attention	0.87	0.86	0.95	0.90	0.93

7.6 Conclusions

This chapter's results demonstrate that the trend-aware attention method outperforms both the baseline and heuristic approaches, achieving the highest accuracy, recall, F1, and AUC. By effectively capturing temporal trends and emphasizing informative frames, the trend-aware method provides a robust and accurate solution for video-level probability aggregation.

A key advantage of the trend-aware approach is its ability to selectively emphasize informative frames. While peak-weighted methods prioritize moments of maximum intensity, they may overlook valuable contextual information, especially when engagement develops gradually rather than manifest in sudden bursts. The trend-aware method mitigates this limitation by considering both the intensity and trajectory of engagement over time. This ensures that the final video-level probability estimation is not only more stable but also more reflective of the user's overall experience throughout the VR interaction.

The implications of these findings extend beyond engagement classification in VR. The superior performance of the trend-aware attention method underscores the importance of temporal modeling in deep learning applications for engagement analysis. Traditional frame-by-frame classification often fails to account for the dynamic nature of interactive VR environments, where user interest fluctuates over time. This study highlights the need for more sophisticated temporal-aware models that can track engagement in real-time.

For VR-based e-commerce, businesses can leverage this approach to accurately evaluate customer interest during virtual interactions. By identifying patterns of engagement and disengagement, companies can personalize user experiences, optimize virtual store layouts, and incorporate interactive features to maximize user retention. A deeper understanding of how users engage with VR shopping environments can also inform decisions about product placement, promotional strategies, and overall UX design.

The impact of this research extends to a broader range of applications beyond VR shopping. The ability of the trend-aware attention method to identify and analyze engagement trends has potential use cases in education, gaming, and telemedicine. In virtual learning environments, educators could detect shifts in student attention and adjust lesson delivery to maintain engagement. In telehealth, emotion and engagement tracking could enhance remote patient interactions, helping healthcare providers assess patient responsiveness in real-time. In gaming, adaptive experiences could be designed based on real-time emotional cues, making virtual environments more immersive and responsive.

Given the effectiveness of this method, future research could build upon these findings by incorporating multimodal data. Combining facial expression recognition with physiological and behavioral signals, such as eye tracking, voice modulation, and heart rate variability, could further enhance engagement prediction accuracy. Additionally, integrating more advanced deep learning architectures, such as transformers or recurrent neural networks, could improve temporal modeling by capturing long-term dependencies in engagement trends. Expanding these methodologies into real-time applications would allow businesses and educators to dynamically adapt experiences based on live engagement feedback and treat more personalized and responsive virtual interactions. As AI-driven emotion and behavior analysis continues to evolve, approaches like this will be instrumental in shaping the future of personalized and interactive VR experiences across industries, from e-commerce to education, healthcare, and entertainment.

Chapter 8

Conclusions

8.1 Research Contributions

This dissertation has made several significant contributions to the fields of artificial intelligence (AI) and e-commerce using a systems engineering approach. These research contributions focus on advancing methodologies and frameworks that integrate facial emotion recognition, generative AI, and virtual reality. These contributions address critical gaps in the existing literature and provide actionable understanding for both academic and practical applications.

This research has made significant advancements in FER methodologies by optimizing and validating convolutional neural network architectures (Xception and ResNet-50) to enhance the accuracy and generalizability of FER models in business research. By leveraging diverse datasets including synthetic data generated through generative adversarial networks, this dissertation has improved the ability of FER systems to classify emotional responses across a wide range of demographics and scenarios. In addition to that, the integration of generative AI techniques has addressed the inherent dataset limitations often encountered in FER and vision deep learning research [236]. The use of synthetic data in combination with real-world datasets sourced from social media has demonstrated the potential to improve model robustness and mitigate overfitting and enhance the applicability of FER models in diverse e-commerce applications.

Moreover, this work has explored the transformative role of immersive virtual reality environments in shaping consumer behavior and user experiences. By examining VR design elements, cognitive load, and exposure sequencing, the research highlights their influence on engagement and purchase intent, and the integration of FER into VR environments has provided new findings into the emotional dynamics of VR based shopping experiences, which offers a deeper understanding of how consumers interact with immersive technologies.

This dissertation has contributed to the development of ethical AI practices within business applications by addressing critical challenges related to data privacy and algorithmic bias as well as transparency. The proposed frameworks emphasize fairness, inclusivity, and user trust, thereby establishing a foundation for more sustainable and ethically responsible AI-driven strategies [237]. Collectively, these contributions advance the intersection of AI, consumer behavior, and business research using a systems engineering approach to provide a roadmap for future investigations.

8.2 Revisiting Research Aims and Research Questions

This dissertation addressed four primary research aims, each associated with specific research questions. Below is a summary of the aims, research questions, and the key findings, in brief, derived from the results.

8.2.1 Aim 1 Findings

Effectiveness of CNNs in Predicting Consumer Engagement

RQ 1.1: How effective are two prominent convolutional neural networks (CNN) architectures, Xception and ResNet-50, in distinguishing consumer engagement (interested vs. disinterested)?

Answer: Both Xception and ResNet-50 demonstrated strong predictive capabilities by high accuracy and F1 score in predicting consumer engagement, with Xception achieving slightly better performance due to its use of depthwise separable convolutions.

RQ 1.2: What role do specific emotional cues (e.g., happiness, disgust, fear, anger, etc.) play in consumer interest classification?

Answer: Positive emotions such as happiness and surprise strongly correlated with interest, while disgust and fear were key indicators of disinterest.

RQ 1.3: What are the implications of facial expression analysis findings for personalized marketing strategies in digital advertising?

Answer: Recognizing emotion-based engagement enables highly targeted and personalized marketing and improves the relevance of advertisements to consumers. The results suggest that

integrating FER with AI-driven recommendation systems can optimize the relevance of advertisements to consumers.

8.2.2 Aim 2 Findings

Addressing Dataset Limitations with Generative AI and Social Media Data

RQ 2.1: Can FER model generalizability be improved using data extracted from social media (YouTube) and/or generated using AI (GANs), as compared to controlled data from a laboratory study?

Answer: Synthetic data generated using Generative Adversarial Networks (GANs), specifically StyleGAN2, enriched underrepresented emotional categories, which led to significant performance improvements in the FER model. Similarly, the inclusion of YouTube data increased generalizability, allowing the model to better handle unseen data, as seen through higher precision-recall values.

RQ 2.2: How can FER models trained on specific categories of advertisements be generalizable to new categories of advertisements?

Answer: The model was able to perform with high accuracy when validated and tested on reaction videos towards advertisements not seen in the training data, suggesting that the model was agnostic of the type of advertisements used in training.

RQ 2.3: What are the ethical and practical considerations in using synthetic and real-world data for FER model training?

Answer: Key concerns include ensuring transparency, avoiding bias propagation, and securing user consent for real-world data.

8.2.3 Aim 3 Findings

Effects of VR Design Complexity and Exposure Sequencing on Engagement

RQ 3.1: How do individual differences, such as technological savviness, brand familiarity, and frequency of online gaming and shopping, moderate the relationship between VR complexity and engagement?

Answer: Greater brand familiarity led to higher immersion levels, while tech-savviness and frequency of playing computer games and shopping online had no effect on engagement.

RQ 3.2: How does designed visual complexity and opportunities for engagement within VR affect perceived immersion, engagement, realism, sense of presence, distraction, effort, and purchase intent?

Answer: More interactive and visually complex VR environments significantly enhanced engagement and purchase intent.

RQ 3.3: How do perceptions of VR interactions, such as perceived immersion, engagement, realism, etc., influence likelihood to purchase?

Answer: Perceived engagement, realism, and effort increased purchase intent, indicating that the increased complexity of the environment did not overwhelm the participants.

RQ 3.4: What are the cognitive and emotional effects of exposure sequencing from a simpler VR to a more detailed VR environment and vice-versa?

Answer: First-exposure bias played a role when participants experienced a simpler VR environment first, where they reported stronger novelty effects. Users transitioning from simpler to complex VR environments reported higher engagement and perceived value compared to those who started with high-complexity environments where cognitive overload was more prevalent.

RQ 3.5: How does cognitive load interact with VR complexity in decision-making?

Answer: While higher complexity increased cognitive load, this was mitigated by user familiarity and interactive guidance, leading to an increased likelihood to purchase and increased perceived engagement and realism.

8.2.4 Aim 4 Findings

Extending Vision Deep Learning Models to VR

RQ 4.1: How does accuracy compare for frame-level versus temporal-level CNN-based FER models in classifying user interest/disinterest in interactive VR environments?

Answer: Frame-level classification, where each frame is treated as an independent observation, fails to capture the sequential nature of engagement. In contrast, the trend-aware attention method achieved the highest accuracy, outperforming traditional methods such as mean, median, Gaussian-weighted, and peak-weighted aggregation. Empirical results show that the trend-aware method achieved the highest accuracy, recall, F1-score, and AUC, demonstrating its ability to distinguish between interested and not-interested users with greater reliability.

8.3 Limitations

This research, while advancing the integration of facial emotion recognition with virtual reality applications, acknowledges several limitations that need further exploration. One important limitation is the diversity of datasets. Although synthetic data was used to enhance variability, the real-world datasets used in this study may still underrepresent certain demographics and cultural expressions that potentially affect the generalizability of FER models.

Additionally, the computational demands of deploying resource-intensive models such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) impose constraints. These requirements could limit the scalability of the proposed methodologies, particularly for smaller organizations or applications requiring real-time processing.

Furthermore, while this research proposed multi-modal frameworks, its primary focus remained on the integration of FER and virtual reality (VR). As such, there is an opportunity for future research to include a more comprehensive analysis that incorporates physiological and behavioral data for deeper findings.

Another limitation arises from the controlled experimental labs, which, while ensuring consistency, as mentioned above may not fully capture the variability and unpredictability of real-world e-commerce scenarios. Specifically, the study conducted as part of this dissertation for Research Aims 3 and 4 poses potential limitations that should be noted. The first one is the sample size (N=55) which is relatively small and predominantly comprised of undergraduate students from a university in California, which may limit the generalizability of the findings to broader populations

with varying demographic and cultural characteristics. Future studies should address this limitation by considering larger and more diverse participant samples to ensure robust and widely applicable conclusions. Additionally, the study relied on self-reported measures for evaluating variables such as engagement, immersion, and purchase intent. Although these metrics are widely used in behavioral research, they are inherently subjective and susceptible to response biases [238].

Integrating objective measures such as eye tracking [239], facial emotion recognition [4] and physiological data [240] in future research could complement self-reported data and provide a more comprehensive understanding of user interactions with VR environments. Also, the focus on two brands (Tesla and Apple) may have influenced participants' responses due to preexisting brand perceptions and loyalties. Future work could expand the scope of VR environments to include a wider range of products and brands, thereby examining whether the observed effects are consistent across various contexts.

Lastly, ethical and privacy considerations, though addressed in this research, need further investigation into user perspectives on data privacy and consent; especially as AI technologies continue to evolve. These limitations emphasize the need for ongoing research to build upon the foundations developed in this dissertation.

8.4 Future Work

This dissertation provides a foundation for advancing the integration of AI and VR technologies in business. However, several opportunities for further exploration remain.

One critical area for future work is enhancing the diversity of datasets used in FER research. While this dissertation utilized synthetic data to address some diversity challenges, real-world datasets often remain underrepresentative of the wide range of demographics and cultural conditions encountered in practice [241]. To ensure FER models can generalize effectively, future research must prioritize the creation and utilization of datasets that capture this diversity [31]. Collaborations with global organizations and cross-cultural studies could facilitate access to such datasets that enable FER systems to better reflect the variability and richness of human emotion.

Another notable direction involves the development of real-time FER and VR applications [242]. Practical deployment in dynamic environments such as e-commerce may require FER models that are both lightweight and computationally efficient [243]. These models must be capable of delivering rapid predictions without compromising accuracy to ensure accurate integration into VR systems. Currently, optimizing VR environments for real-time interaction is essential and this includes reducing latency to ensure smooth transitions between virtual experiences [244]. Achieving real-time FER and VR integration will unlock transformative applications such as adaptive virtual shopping experiences that respond instantly to consumer emotions [245].

Expanding the scope of multi-modal frameworks is another promising field of study for future research. While this dissertation focused primarily on FER, incorporating additional data modalities, such as physiological signals like heart rate, skin conductance, and eye-tracking data, offers the potential to provide a richer understanding of consumer behavior [246–250]. These data sources could complement FER by capturing emotional and cognitive states that are not readily observable through facial expressions alone and by doing that we can make comprehensive inferences into consumer decision-making processes and enhance personalization and the overall effectiveness of business strategies [244].

Current research primarily examines immediate outcomes such as engagement and purchase intent. However, understanding the long-term effects of VR based shopping experiences on consumer behavior is equally important [81]. Future studies should investigate how repeated interactions with VR environments influence factors such as customer satisfaction and retention over extended periods. This longitudinal perspective will provide a more nuanced understanding of VR's impact on consumer behavior and inform the design of sustainable virtual experiences.

Ethical and policy considerations remain a vital area for future exploration. The use of FER and VR in business raises significant ethical questions, including data privacy, consent, and algorithmic fairness [251]. Further research is needed to develop robust ethical guidelines and policy frameworks that address these concerns. These frameworks must ensure transparency in data col-

lection and use established mechanisms for accountability and prioritize fairness in algorithmic decision making.

Lastly, the methodologies developed in this dissertation can be incorporated into applications beyond business and e-commerce. The generalizability of FER and VR frameworks to other industries such as healthcare, education, entertainment, and many others represents a valuable direction for future research. For instance, FER could be used to monitor patient well-being in healthcare, while VR could transform educational experiences by creating immersive learning environments [252] By extending these technologies to diverse domains, researchers can unlock their transformative potential to address a broader range of societal challenges.

8.5 Publications of Results

The findings and methodologies presented in this dissertation have led to the following peer-reviewed publications. These journal papers contribute to the academic and professional fields of systems engineering, AI, and business.

1. Alipour, P., Gallegos, E.E., & Sridhar, S. (2024). AI-driven marketing personalization: Deploying convolutional neural networks to decode consumer behavior. *International Journal of Human–Computer Interaction*, 1–19. doi.org/10.1080/10447318.2024.2432455

    - Journal Impact Factor: 3.4

2. Alipour, P., & Gallegos, E.E. (2025). Leveraging generative AI synthetic and social media data for content generalizability to overcome data constraints in vision deep learning. *Journal of Artificial Intelligence Review*. doi.org/10.1007/s10462-025-11137-6

    - Journal Impact Factor: 10.7

3. Alipour, P., Gallegos, E.E., & Sridhar, S.(submitted for review). Multi-modal causal analysis of product exploration: Examining impacts of immersion and exposure sequences on consumer behavior in virtual reality e-commerce. Under review at *Journal of Business Research*.

- Journal Impact Factor: 10.5

4. Alipour, P., & Gallegos, E.E. (forthcoming). AI-powered virtual reality: Enhancing user experience in VR e-commerce through facial emotion recognition. Target journal: *Journal of Virtual Reality*.

    - Journal Impact Factor: 4.4

Bibliography

[1] Xi Li, Mengze Shi, and Xin Shane Wang. Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing*, 36(2):216–231, 2019.

[2] Stefano Puntoni, Rebecca Walker Reczek, Markus Giesler, and Simona Botti. Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, 85(1):131–151, 2021.

[3] Liu Liu, Daria Dzyabura, and Natalie Mizik. Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4):669–686, 2020.

[4] Panteha Alipour, Erika E Gallegos, and Shrihari Sridhar. AI-driven marketing personalization: Deploying convolutional neural networks to decode consumer behavior. *International Journal of Human–Computer Interaction*, pages 1–19, 2024.

[5] Yoesoep Edhie Rachmad. *Digital Marketing Theories: From Gimmicks to Loyalty*. PT. Sonpedia Publishing Indonesia, 2024.

[6] Yawen Li. Optimizing sentiment analysis of user reviews and emotional marketing strategies on e-commerce platforms using deep learning algorithms. *Journal of Electrical Systems*, 20(9s):437–444, 2024.

[7] Mi Zhou, George H Chen, Pedro Ferreira, and Michael D Smith. Consumer behavior in the online classroom: Using video analytics and machine learning to understand the consumption of video courseware. *Journal of Marketing Research*, 58(6):1079–1100, 2021.

[8] Anirudh Sai Vallabhaneni, Anjali Perla, Revanth Reddy Regalla, and Neelam Kumari. The power of personalization: Ai-driven recommendations. In *Minds Unveiled*, pages 111–127. Productivity Press, 2024.

[9] Patrick Azuka Okeleke, Daniel Ajiga, Samuel Olaoluwa Folorunsho, and Chinedu Ezeigweneme. Predictive analytics for market trends using ai: A study in consumer behavior. *International Journal of Engineering Research Updates*, 7(1):36–49, 2024.

[10] Pranali Dhawas, Aparna Bondade, Sandhya Patil, Kiran Shyam Khandare, and Ramadevi V Salunkhe. Intelligent automation in marketing. In *Hyperautomation in Business and Society*, pages 66–88. IGI Global, 2024.

[11] Saif Ahmed and Norzalita Abd Aziz. Impact of ai on customer experience in video streaming services: A focus on personalization and trust. *International Journal of Human–Computer Interaction*, 39:1–12, 2024.

[12] Farah Saboune. Ai-driven marketing strategies: Unlocking growth potential and operational efficiency in the digital communication landscape. In *2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*, pages 295–301. IEEE, 2024.

[13] Ahmed Nassar and Mostafa Kamal. Ethical dilemmas in ai-powered decision-making: a deep dive into big data-driven ethical considerations. *International Journal of Responsible Artificial Intelligence*, 11(8):1–11, 2021.

[14] Shuili Du and Chunyan Xie. Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research*, 129:961–974, 2021.

[15] Prathyusha Nama, Suprit Pattanayak, and Harika Sree Meka. Ai-driven innovations in cloud computing: Transforming scalability, resource management, and predictive analytics in distributed systems. *International Research Journal of Modernization in Engineering Technology and Science*, 5(12):4165, 2023.

[16] Jan R Landwehr, Ann L McGill, and Andreas Herrmann. It's got the look: The effect of friendly and aggressive "facial" expressions on product liking and sales. *Journal of marketing*, 75(3):132–146, 2011.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Yanhao "Max" Wei, Jihoon Hong, and Gerard J Tellis. Machine learning for creativity: Using similarity networks to design better crowdfunding projects. *Journal of Marketing*, 86(2):87–104, 2022.

[19] Robert V Kozinets and Ulrike Gretzel. Commentary: Artificial intelligence: The marketer's dilemma. *Journal of Marketing*, 85(1):156–159, 2021.

[20] Yoren Gaffary, Victoria Eyharabide, Jean-Claude Martin, and Mehdi Ammi. The impact of combining kinesthetic and facial expression displays on emotion recognition by users. *International Journal of Human-Computer Interaction*, 30(11):904–920, 2014.

[21] Xinyi Huang and Daniela M Romano. Coral morph: An artistic shape-changing textile installation for mindful emotion regulation in the wild. *International Journal of Human–Computer Interaction*, pages 1–17, 2024.

[22] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187, 2003.

[23] Sulis Sandiwarno, Zhendong Niu, and Ally S Nyamawe. Ses-net: A novel multi-task deep neural network model for analyzing e-learning users' satisfaction via sentiment, emotion, and semantic. *International Journal of Human–Computer Interaction*, pages 1–24, 2024.

[24] Clarice O'Brien and Styliani Vlachou. Facial expressions and recognition for the communication of thoughts and emotions in business and marketing. *Biometrics and Neuroscience Research in Business and Management*, page 99, 2021.

[25] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[26] Iris Dominguez-Catena, Daniel Paternain, Aranzazu Jurio, and Mikel Galar. Less can be more: representational vs. stereotypical gender bias in facial expression recognition. *Progress in Artificial Intelligence*, pages 1–21, 2024.

[27] Panteha Alipour and Erika Gallegos. Leveraging generative ai synthetic and social media data for content generalizability to overcome data constraints in vision deep learning. *Artificial Intelligence Review*, 58(5), 2025.

[28] Martina Mattioli and Federico Cabitza. Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. *Machine Learning and Knowledge Extraction*, 6(4):2201–2231, 2024.

[29] George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. Human-centered design of artificial intelligence. In Gavriel Salvendy and Waldemar Karwowski, editors, *Handbook of Human Factors and Ergonomics*, chapter 42, pages 1–20. Wiley, 2021.

[30] Amira Mouakher and Ruslan Kononov. Explainable evaluation framework for facial expression recognition in web-based learning environments. *International Journal of Machine Learning and Cybernetics*, pages 1–33, 2024.

[31] Thomas Kopalidis, Vassilios Solachidis, Nicholas Vretos, and Petros Daras. Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information*, 15(3):135, 2024.

[32] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors*, 19(8):1863, 2019.

[33] Christian M Derbaix. The impact of affective reactions on attitudes toward the advertisement and the brand: A step toward ecological validity. *Journal of marketing research*, 32(4):470–479, 1995.

[34] Josephine LCM Woltman Elpers, Michel Wedel, and Rik GM Pieters. Why do consumers stop viewing television commercials? two experiments on the influence of moment-to-moment entertainment and information value. *Journal of Marketing Research (JMR)*, 40(4), 2003.

[35] Yinghui Zhou, Shasha Lu, and Min Ding. Contour-as-face framework: A method to preserve privacy and perception. *Journal of Marketing Research*, 57(4):617–639, 2020.

[36] Demetrios Vakratsas and Tim Ambler. How advertising works: what do we really know? *Journal of marketing*, 63(1):26–43, 1999.

[37] Amr E Eldin Rashed, Ahmed E Mansour Atwa, Ali Ahmed, Mahmoud Badawy, Mostafa A Elhosseini, and Waleed M Bahgat. Facial image analysis for automated suicide risk detection with deep neural networks. *Artificial Intelligence Review*, 57(10):1–43, 2024.

[38] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9):607–619, 2011.

[39] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[40] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[41] Zhe Sun, Jiatong Bai, Panpan Wang, and Jiaxue Huang. Combining deep subspace feature representation based ikpcanet and jointly constraint multi-dictionary learning for facial expression recognition. *Artificial Intelligence Review*, 56(Suppl 1):937–958, 2023.

[42] Alan S Cowen, Dacher Keltner, Florian Schroff, Brendan Jou, Hartwig Adam, and Gautam Prasad. Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841):251–257, 2021.

[43] Tulika Chutia and Nomi Baruah. A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, 57(8):203, 2024.

[44] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[45] Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, 55(3):2351–2377, 2022.

[46] Teerath Kumar, Rob Brennan, Alessandra Mileo, and Malika Bendechache. Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*, 2024.

[47] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020.

[48] Haythem Ghazouani. Challenges and emerging trends for machine reading of the mind from facial expressions. *SN Computer Science*, 5(1):103, 2023.

[49] Sepideh Kalateh, Luis A Estrada-Jimenez, Sanaz Nikghadam Hojjati, and Jose Barata. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. *IEEE Access*, 2024.

[50] David Kornish, Soundararajan Ezekiel, and Maria Cornacchia. Dcnn augmentation via synthetic data from variational autoencoders and generative adversarial networks. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6. IEEE, 2018.

[51] Yaganteeswarudu Akkem, Saroj Kumar Biswas, and Aruna Varanasi. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. *Engineering Applications of Artificial Intelligence*, 131:107881, 2024.

[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[53] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[54] Vignesh Sampath, Iñaki Maurtua, Juan Jose Aguilar Martin, and Aitor Gutierrez. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of big Data*, 8:1–59, 2021.

[55] Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. A survey of unsupervised generative models for exploratory data analysis and representation learning. *Acm computing surveys (csur)*, 54(5):1–40, 2021.

[56] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018.

[57] Christian Mejia-Escobar, Miguel Cazorla, and Ester Martinez-Martin. Improving facial expression recognition through data preparation & merging. *IEEE Access*, 2023.

[58] Gianmarco Ipinze Tutuianu, Yang Liu, Ari Alamäki, and Janne Kauttonen. Benchmarking deep facial expression recognition: An extensive protocol with balanced dataset in the wild. *Engineering Applications of Artificial Intelligence*, 136:108983, 2024.

[59] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

[60] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, Li Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

[61] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022.

[62] Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[63] Lacramioara Mazilu, Norman W Paton, Nikolaos Konstantinou, and Alvaro AA Fernandes. Fairness-aware data integration. *ACM Journal of Data and Information Quality*, 14(4):1–26, 2022.

[64] Marcus AK September, Francesco Sanna Passino, Leonie Goldmann, and Anton Hinel. Extended deep adaptive input normalization for preprocessing time series data for neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1891–1899. PMLR, 2024.

[65] Zahid ur Rahman, Mohd Shahrimie Mohd Asaari, Haidi Ibrahim, Intan Sorfina Zainal Abidin, and Mohamad Khairi Ishak. Generative adversarial networks (gans) for image augmentation in farming: A review. *IEEE Access*, 2024.

[66] Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makaravets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. Face generation and editing with stylegan: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[67] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022.

[68] Aadith Sukumar, Aditya Desai, Peeyush Singhal, Sai Gokhale, Deepak Kumar Jain, Rahee Walambe, and Ketan Kotecha. Training against disguises: Addressing and mitigating bias in facial emotion recognition with synthetic data. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2024.

[69] Amit Kumar. The role of synthetic data in advancing ai models: Opportunities, challenges, and ethical considerations. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5(1):443–459, 2024.

[70] Sheikh Inam Ul Mansoor. Legal implications of deepfake technology: In the context of manipulation, privacy, and identity theft. *Central University of Kashmir Law Review*, 4:65–92, 2024.

[71] Philippe De Wilde, Payal Arora, Fernando Buarque de Lima Neto, Yik Chin, Mamello Thinyane, Serge Stinckwich, Eleonore Fournier-Tombs, and T Marwala. Recommendations on the use of synthetic data to train ai models. *United Nations University*, 2024.

[72] Geetika Madaan, Satish Kumar Asthana, and Jaskiran Kaur. Generative ai: Applications, models, challenges, opportunities, and future directions. *Generative AI and Implications for Ethics, Security, and Data Management*, pages 88–121, 2024.

[73] Ritesh Raj, Kuldeep Kumar, Aditya Prakash, Abhishek Kumar, Sujata Kumari, and Guddu Kumar. Enhancing e-commerce engagement: Exploring ar and vr-based marketing strate-

gies. In *2024 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, volume 1, pages 448–453. IEEE, 2024.

[74] Qian Janice Wang, Francisco Barbosa Escobar, Patricia Alves Da Mota, and Carlos Velasco. Getting started with virtual reality for sensory and consumer science: Current practices and future perspectives. *Food Research International*, 145:110410, 2021.

[75] Costas Boletsis and Amela Karahasanovic. Immersive technologies in retail: Practices of augmented and virtual reality. In *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications*. SciTePress, 2020.

[76] D Eric Boyd and Bernadett Koles. Virtual reality and its impact on b2b marketing: A value-in-use perspective. *Journal of Business Research*, 100:590–598, 2019.

[77] Sandra Maria Correia Loureiro, João Guerreiro, Sara Eloy, Daniela Langaro, and Padma Panchapakesan. Understanding the use of virtual reality in marketing: A text mining-based review. *Journal of Business Research*, 100:514–530, 2019.

[78] Kun Chang Lee and Namho Chung. Empirical analysis of consumer reaction to the virtual reality shopping mall. *Computers in Human Behavior*, 24(1):88–104, 2008.

[79] Zhenhui Jiang and Izak Benbasat. Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping. *Journal of Management Information Systems*, 21(3):111–147, 2004.

[80] Helena Van Kerrebroeck, Malaika Brengman, and Kim Willems. When brands come to life: experimental research on the vividness effect of virtual reality in transformational marketing communications. *Virtual Reality*, 21:177–191, 2017.

[81] Gabriele Pizzi, Daniele Scarpi, Marco Pichierri, and Virginia Vannucci. Virtual reality, real reactions?: Comparing consumers' perceptions and shopping orientation across physical and virtual-reality retail stores. *Computers in Human Behavior*, 96:1–12, 2019.

[82] Hyun Jung Oh, Junghwan Kim, Jeongheon JC Chang, Nohil Park, and Sangrock Lee. Social benefits of living in the metaverse: The relationships among social presence, supportive interaction, social self-efficacy, and feelings of loneliness. *Computers in Human Behavior*, 139:107498, 2023.

[83] Mel Slater and Martin Usoh. Presence in immersive virtual environments. In *Proceedings of IEEE virtual reality annual international symposium*, pages 90–96. IEEE, 1993.

[84] Lina Leimontaitė and Jurga Naimavičienė. Virtual 3d tour assistance in real estate management. *Baltic Journal of Real Estate Economics and Construction Management*, 11(1):153–159, 2023.

[85] Mohamad Zaidi Sulaiman, Mohd Nasiruddin Abdul Aziz, Mohd Haidar Abu Bakar, Nur Akma Halili, and Muhammad Asri Azuddin. Matterport: virtual tour as a new marketing approach in real estate business during pandemic covid-19. In *International conference of innovation in media and visual design (IMDES 2020)*, pages 221–226. Atlantis Press, 2020.

[86] Jesus Martínez-Navarro, Enrique Bigné, Jaime Guixeres, Mariano Alcañiz, and Carmen Torrecilla. The influence of virtual reality in e-commerce. *Journal of Business Research*, 100:475–482, 2019.

[87] Kamal Upreti, Divya Gangwar, Prashant Vats, Rishu Bhardwaj, Vishal Khatri, and Vijay Gautam. Artificial neural networks for enhancing e-commerce: A study on improving personalization, recommendation, and customer experience. In *International Conference on Electrical and Electronics Engineering*, pages 141–153. Springer, 2023.

[88] Aishwarya Gowda AG, Hui-Kai Su, and Wen-Kai Kuo. Personalized e-commerce: Enhancing customer experience through machine learning-driven personalization. In *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, pages 1–5. IEEE, 2024.

[89] Jonathan Liebers, Sascha Brockel, Uwe Gruenefeld, and Stefan Schneegass. Identifying users by their hand tracking data in augmented and virtual reality. *International Journal of Human–Computer Interaction*, 40(2):409–424, 2024.

[90] Michel Wedel, Enrique Bigné, and Jie Zhang. Virtual and augmented reality: Advancing research in consumer marketing. *International Journal of Research in Marketing*, 37(3):443–465, 2020.

[91] Rubén Grande, Javier Albusac, David Vallejo, Carlos Glez-Morcillo, and José Jesús Castro-Schez. Performance evaluation and optimization of 3d models from low-cost 3d scanning technologies for virtual reality and metaverse e-commerce. *Applied Sciences*, 14(14):6037, 2024.

[92] Hendrik Büchel and Stefan Spinler. The impact of the metaverse on e-commerce business models–a delphi-based scenario study. *Technology in Society*, 76:102465, 2024.

[93] Rilliandi Arindra Putawa and Dwi Sugianto. Exploring user experience and immersion levels in virtual reality: A comprehensive analysis of factors and trends. *International Journal Research on Metaverese*, 1(1):20–39, 2024.

[94] George Caleb Oguta. Securing the virtual marketplace: Navigating the landscape of security and privacy challenges in e-commerce. *GSC Advanced Research and Reviews*, 18(1):084–117, 2024.

[95] Oladri Renuka, Niranchana RadhaKrishnan, Bodapatla Sindhu Priya, Avula Jhansy, and Soundarajan Ezekiel. Data privacy and protection: Legal and ethical challenges. *Emerging Threats and Countermeasures in Cybersecurity*, pages 433–465, 2025.

[96] Yulin Chen. Investigating cross-cultural generalizability of facial emotion recognition with multi-dataset training. B.S. thesis, University of Twente, 2024.

[97] Alex Boutin, Lucie Lévêque, and Sonia Desmoulin-Canselier. On legal and ethical challenges of automatic facial expression recognition: An exploratory study. In *Proceedings of*

the 2023 ACM International Conference on Interactive Media Experiences, pages 226–229, 2023.

[98] Jyothi Pinnika. Evaluating privacy and security concerns of facial emotion recognition system. Master's thesis, Southeast Missouri State University, 2020.

[99] Danish Ali, Sundas Iqbal, Shahid Mehmood, Irshad Khalil, and Inam Ullah. Advances, challenges, and collaborative. *Artificial General Intelligence (AGI) Security: Smart Applications and Sustainable Technologies*, page 211, 2024.

[100] Siwei Lyu. Deepfake the menace: mitigating the negative impacts of ai-generated content. *Organizational Cybersecurity Journal: Practice, Process and People*, 2024.

[101] Chandrasekhar Uddagiri and Bala Venkateswarlu Isunuri. Ethical and privacy challenges of generative ai. In *Generative AI: Current Trends and Applications*, pages 219–244. Springer, 2024.

[102] Nawaf Waqas, Sairul Izwan Safie, Kushsairy Abdul Kadir, Sheroz Khan, and Muhammad Haris Kaka Khel. Deepfake image synthesis for data augmentation. *IEEE Access*, 10:80847–80857, 2022.

[103] Deepak Kaul and Rahul Khurana. Ai-driven optimization models for e-commerce supply chain operations: Demand prediction, inventory management, and delivery time reduction with cost efficiency considerations. *International Journal of Social Analytics*, 7(12):59–77, 2022.

[104] Muhammad Sohail, Ghulam Ali, Javed Rashid, Israr Ahmad, Sultan H Almotiri, Mohammed A AlGhamdi, Arfan A Nagra, and Khalid Masood. Racial identity-aware facial expression recognition using deep convolutional neural networks. *Applied Sciences*, 12(1):88, 2021.

[105] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[106] Jeff JH Kim, Adith V Srivatsa, George R Nahass, Timur Rusanov, Soonmyung Hwang, Soohyun Kim, Itay Solomon, Tae Ha Lee, Shrinidhi Kadkol, Olusola Ajilore, et al. Generative ai can effectively manipulate data. *AI and Ethics*, pages 1–15, 2024.

[107] P Kumar, K Deivanai, S Srivathsav, M Uthandeeswar, and S Senthil Pandi. A novel approach for face generator based on emotions. In *2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST)*, pages 1–6. IEEE, 2024.

[108] Gustav Bøg Petersen, Giorgos Petkakis, and Guido Makransky. A study of how immersion and interactivity drive vr learning. *Computers & Education*, 179:104429, 2022.

[109] Jengchung Victor Chen, Quang-An Ha, and Minh Tam Vu. The influences of virtual reality shopping characteristics on consumers' impulse buying behavior. *International Journal of Human–Computer Interaction*, 39(17):3473–3491, 2023.

[110] Nannan Xi and Juho Hamari. Shopping in virtual reality: A literature review and future agenda. *Journal of Business Research*, 134:37–58, 2021.

[111] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. Multi-modal preference modeling for product search. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1865–1873, 2018.

[112] Büşra Kocaçınar, Pelin İnan, Ela Nur Zamur, Buket Çalşimşek, Fatma Patlar Akbulut, and Cagatay Catal. Neurobiosense: A multidimensional dataset for neuromarketing analysis. *Data in Brief*, page 110235, 2024.

[113] Rahul Sharma, Shramishtha Srivastva, and Sanobar Fatima. E-commerce and digital transformation: Trends, challenges, and implications. *International Journal for Multidisciplinary Research (IJFMR)*, 5(5), 2023.

[114] PT Williams. Emerging trends in e-commerce: A global perspective. *International Journal of Open Publication and Exploration, ISSN: 3006-2853*, 7(1):25–30, 2019.

[115] Ju Yeun Jang, Eunsoo Baek, So-Yeon Yoon, and Ho Jung Choo. Store design: Visual complexity and consumer responses. *International Journal of Design*, 12(2), 2018.

[116] Shengliang Zhang, Guanyu Tang, Xiaodong Li, and Ai Ren. The effects of appearance personification of service robots on customer decision-making in the product recommendation context. *Industrial Management & Data Systems*, 123(2):578–595, 2023.

[117] Mritunjay Rai and Jay Kumar Pandey. *Using Machine Learning to Detect Emotions and Predict Human Psychology*. IGI Global, 2024.

[118] Baba Shiv and Alexander Fedorikhin. Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of consumer Research*, 26(3):278–292, 1999.

[119] Albert Mehrabian. *Silent messages*, volume 8. Wadsworth Belmont, CA, 1972.

[120] Andre Teixeira Lopes, Edilson De Aguiar, and Thiago Oliveira-Santos. A facial expression recognition system using convolutional networks. In *2015 28th SIBGRAPI conference on graphics, patterns and images*, pages 273–280. IEEE, 2015.

[121] Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, page 102019, 2023.

[122] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[123] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 2019.

[124] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[125] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[126] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[127] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[128] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[129] Reagan E Mandiya, Hervé M Kongo, Selain K Kasereka, Kyamakya Kyandoghere, Petro Mushidi Tshakwanda, and Nathanaël M Kasoro. Enhancing covid-19 detection: An xception-based model with advanced transfer learning from x-ray thorax images. *Journal of Imaging*, 10(3):63, 2024.

[130] Kashif Shaheed, Aihua Mao, Imran Qureshi, Munish Kumar, Sumaira Hussain, Inam Ullah, and Xingming Zhang. Ds-cnn: A pre-trained xception model based on depth-wise separable convolutional neural network for finger vein recognition. *Expert Systems with Applications*, 191:116288, 2022.

[131] CS Anumol. Advancements in cnn architectures for computer vision: A comprehensive review. In *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, pages 1–7. IEEE, 2023.

[132] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[133] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[134] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[135] Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *international conference on machine learning*, pages 2217–2225. PMLR, 2016.

[136] Hidenori Ide and Takio Kurita. Improvement of learning for cnn with relu activation by sparse regularization. In *2017 international joint conference on neural networks (IJCNN)*, pages 2684–2691. IEEE, 2017.

[137] Sungheon Park and Nojun Kwak. Analysis on the dropout effect in convolutional neural networks. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 189–204. Springer, 2017.

[138] Christian Garbin, Xingquan Zhu, and Oge Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia tools and applications*, 79(19):12777–12815, 2020.

[139] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):1–43, 2024.

[140] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[141] M Msambare. Fer-2013 dataset, 2017. Accessed: 2024-06-09.

[142] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild, 2017. Accessed: 2024-06-09.

[143] Shuvoalok. Ck+ dataset. https://www.kaggle.com/datasets/shuvoalok/ck-dataset, 2021. Accessed: 2024-06-09.

[144] Anas. Facial expression recognition using jaffe dataset. https://github.com/anas-899/facial-expression-recognition-Jaffe, 2023. Accessed: 2024-06-09.

[145] bknyaz. Emotiw dataset. https://github.com/bknyaz/emotiw, 2019. Accessed: 2024-06-09.

[146] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[147] Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in neural information processing systems*, 20, 2007.

[148] Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

[149] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[150] S Saravanan, Hemal Shingloo, Nameera Sajid, Navneet Lamba, Akshyansu Pritam, and Anupam Srivastava. Automatic photo enhancer using machine learning and deep learning

with python. In *International Conference on Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication*, pages 553–563. Springer, 2023.

[151] Wenyi Huang and Jack W Stokes. Mtnet: a multi-task neural network for dynamic malware classification. In *Detection of Intrusions and Malware, and Vulnerability Assessment: 13th International Conference, DIMVA 2016, San Sebastián, Spain, July 7-8, 2016, Proceedings 13*, pages 399–418. Springer, 2016.

[152] Yingjie Tian and Yuqi Zhang. A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, 80:146–166, 2022.

[153] Yaoshiang Ho and Samuel Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*, 8:4806–4813, 2019.

[154] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

[155] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[156] Damir Krstinić, Maja Braović, Ljiljana Šerić, and Dunja Božić-Štulić. Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*, 1:1–14, 2020.

[157] J Terven, DM Cordova-Esparza, A Ramirez-Pedraza, and EA Chavez-Urbiola. Loss functions and metrics in deep learning. *arXiv preprint arXiv:2307.02694*, 2023.

[158] Giulio Biondi, Valentina Franzoni, and Alfredo Milani. Defining classification ambiguity to discover a potential bias applied to emotion recognition data sets. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 672–679. IEEE, 2022.

[159] Florian Bublatzky, Fatih Kavcıoğlu, Pedro Guerra, Sarah Doll, and Markus Junghöfer. Contextual information resolves uncertainty about ambiguous facial emotions: Behavioral and magnetoencephalographic correlates. *NeuroImage*, 215:116814, 2020.

[160] Olufisayo S Ekundayo and Serestina Viriri. Facial expression recognition: A review of trends and techniques. *Ieee Access*, 9:136944–136973, 2021.

[161] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[162] Redhwan Algabri, Ahmed Abdu, and Sungon Lee. Deep learning and machine learning techniques for head pose estimation: a survey. *Artificial Intelligence Review*, 57(10):1–66, 2024.

[163] P Alipour. Facial emotion recognition for content generalizability. https://github.com/pantehaalipourTeslaeng/consumer-fer-content-generalization, 2025.

[164] Donghyuk Shin, Shu He, Gene Moo Lee, Andrew B Whinston, Suleyman Cetintas, and Kuang-Chih Lee. *Enhancing social media analysis with visual data analytics: A deep learning approach*. SSRN Amsterdam, The Netherlands, 2020.

[165] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021.

[166] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11895–11904, 2021.

[167] Haodong Li, Han Chen, Bin Li, and Shunquan Tan. Can forensic detectors identify gan generated images? In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 722–727. IEEE, 2018.

[168] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[169] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. One-shot neural face reenactment via finding directions in gan's latent space. *International Journal of Computer Vision*, pages 1–31, 2024.

[170] Min Jin Chong, Wen-Sheng Chu, Abhishek Kumar, and David Forsyth. Retrieve in style: Unsupervised facial feature transfer and retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3887–3896, 2021.

[171] Tejan Karmali, Rishubh Parihar, Susmit Agrawal, Harsh Rangwani, Varun Jampani, Maneesh Singh, and R Venkatesh Babu. Hierarchical semantic regularization of latent spaces in stylegans. In *European Conference on Computer Vision*, pages 443–459. Springer, 2022.

[172] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12863–12872, 2021.

[173] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[174] Amit H Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Oren Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. In *Computer Graphics Forum*, volume 41, pages 591–611. Wiley Online Library, 2022.

[175] Ran Yuan, Bo Wang, Yeqi Sun, Xuanning Song, and Junzo Watada. Conditional style-based generative adversarial networks for renewable scenario generation. *IEEE Transactions on Power Systems*, 38(2):1281–1296, 2022.

[176] Diederik P Kingma and JL Ba. Adam: A method for stochastic optimization 3rd international conference on learning representations. In *ICLR 2015-Conference Track Proceedings*, volume 1, 2015.

[177] Dominik Lewy and Jacek Mańdziuk. An overview of mixing augmentation methods and augmentation strategies. *Artificial Intelligence Review*, 56(3):2111–2169, 2023.

[178] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[179] Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986, 2020.

[180] Xavier Soria Poma, Edgar Riba, and Angel Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1923–1932, 2020.

[181] Laith Alzubaidi, Mohammed A Fadhel, Freek Hollman, Asma Salhi, Jose Santamaria, Ye Duan, Ashish Gupta, Kenneth Cutbush, Amin Abbosh, and Yuantong Gu. Ssp: self-supervised pertaining technique for classification of shoulder implants in x-ray medical images: a broad experimental study. *Artificial Intelligence Review*, 57(9):261, 2024.

[182] Yixin Liu, Lihang Zhang, Zezhou Hao, Ziyuan Yang, Shanjuan Wang, Xiaoguang Zhou, and Qing Chang. An xception model based on residual attention mechanism for the classification of benign and malignant gastric ulcers. *Scientific Reports*, 12(1):15365, 2022.

[183] Evgin Goceri. Gan based augmentation using a hybrid loss function for dermoscopy images. *Artificial Intelligence Review*, 57(9):234, 2024.

[184] Peng Liu, Wei Qian, Hua Zhang, Yabin Zhu, Qi Hong, Qiang Li, and Yudong Yao. Automatic sleep stage classification using deep learning: signals, data representation, and neural networks. *Artificial Intelligence Review*, 57(11):301, 2024.

[185] Mona Alzahrani, Muhammad Usman, Salma Kammoun Jarraya, Saeed Anwar, and Tarek Helmy. Deep models for multi-view 3d object recognition: a review. *Artificial Intelligence Review*, 57(12):1–71, 2024.

[186] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

[187] Fahim Anzum, Ashratuz Zavin Asha, and Marina L Gavrilova. Biases, fairness, and implications of using ai in social media data mining. In *2022 International Conference on Cyberworlds (CW)*, pages 251–254. IEEE, 2022.

[188] Yingjian Li, Yingnan Gao, Bingzhi Chen, Zheng Zhang, Guangming Lu, and David Zhang. Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):3190–3202, 2021.

[189] Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, and Joel JPC Rodrigues. A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alexandria Engineering Journal*, 68:817–840, 2023.

[190] Miro Mannino and Azza Abouzied. Is this real? generating synthetic data that looks real. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 549–561, 2019.

[191] Johannes Schneider. Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289, 2024.

[192] Derek B Lilienthal. Synthetic data generation for accurate, fair, and private recommender systems. Master's thesis, San Jose State University, 2024.

[193] Declan Humphreys, Abigail Koay, Dennis Desmond, and Erica Mealy. Ai hype as a cyber security risk: the moral responsibility of implementing generative ai in business. *AI and Ethics*, pages 1–14, 2024.

[194] Wyke Stommel and Lynn de Rijk. Ethical approval: None sought. how discourse analysts report ethical issues around publicly available online data. *Research Ethics*, 17(3):275–297, 2021.

[195] Karl Van der Schyff, Stephen Flowerday, and Steven Furnell. Duplicitous social media and data surveillance: An evaluation of privacy risk. *Computers & Security*, 94:101822, 2020.

[196] Arianna Rossi, Mónica P Arenas, Emre Kocyigit, and Moad Hani. Challenges of protecting confidentiality in social media data and their ethical import. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 554–561. IEEE, 2022.

[197] Goran Paulin and Marina Ivasic-Kos. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial intelligence review*, 56(9):9221–9265, 2023.

[198] UHWA Hewage, Roopak Sinha, and M Asif Naeem. Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review. *Artificial Intelligence Review*, 56(9):10427–10464, 2023.

[199] Zhengjing Ma, Gang Mei, and Nengxiong Xu. Generative deep learning for data generation in natural hazard analysis: motivations, advances, challenges, and opportunities. *Artificial Intelligence Review*, 57(6):160, 2024.

[200] Ishfaq Hussain Rather, Sushil Kumar, and Amir H Gandomi. Breaking the data barrier: a review of deep learning techniques for democratizing ai with small datasets. *Artificial Intelligence Review*, 57(9):226, 2024.

[201] Muhammad Atif Butt, Adnan Qayyum, Hassan Ali, Ala Al-Fuqaha, and Junaid Qadir. Towards secure private and trustworthy human-centric embedded machine learning: An emotion-aware facial recognition case study. *Computers & Security*, 125:103058, 2023.

[202] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 638–644, 2021.

[203] Jie Li, Yongli Ren, and Ke Deng. Fairgan: Gans-based fairness-aware learning for recommendations with implicit feedback. In *Proceedings of the ACM web conference 2022*, pages 297–307, 2022.

[204] Moumita Sinha, Yancheng Li, Wei Shung Chung, and Paul Hsiung. Bias correction for supervised learning in email marketing. In *SIGIR eCom'20*, 2020.

[205] Marc Schmitt and Ivan Flechais. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57(12):1–23, 2024.

[206] Belinda Lunnay, Joseph Borlagdan, Darlene McNaughton, and Paul Ward. Ethical use of social media to facilitate qualitative research. *Qualitative health research*, 25(1):99–109, 2015.

[207] Ambika Pawar, Swati Ahirrao, and Prathamesh P Churi. Anonymization techniques for protecting privacy: A survey. In *2018 IEEE Punecon*, pages 1–6. IEEE, 2018.

[208] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in neural information processing systems*, 33:361–371, 2020.

[209] Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

[210] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[211] Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253, 2019.

[212] Attila Dabis and Csaba Csáki. Ai and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative ai. *Humanities and Social Sciences Communications*, 11(1):1–13, 2024.

[213] Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2239–2250, 2022.

[214] Emily C Zabor, Alexander M Kaizer, and Brian P Hobbs. Randomized controlled trials. *Chest*, 158(1):S79–S87, 2020.

[215] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[216] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[217] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[218] Neil Charness and Walter R Boot. Aging and information technology use: Potential and barriers. *Current directions in psychological science*, 18(5):253–258, 2009.

[219] Kerryellen G Vroman, Sajay Arthanat, and Catherine Lysack. "who over 65 is online?" older adults' dispositions toward information communication technology. *Computers in Human Behavior*, 43:156–166, 2015.

[220] Ricardo Lopes and Rafael Bidarra. Adaptivity challenges in games and simulations: a survey. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(2):85–99, 2011.

[221] Muhammad Anas Khalid and Dr Vida Viktoria. Strategic marketing plan for apple inc. *Network Intelligence Studies*, pages 61–74, 2023.

[222] Carlos Flavián, Sergio Ibáñez-Sánchez, and Carlos Orús. The impact of virtual, augmented and mixed reality technologies on the customer experience. *Journal of business research*, 100:547–560, 2019.

[223] Amala V Rajan, Nasser Nassiri, Vishwesh Akre, Rejitha Ravikumar, Amal Nabeel, Maryam Buti, and Fatima Salah. Virtual reality gaming addiction. *2018 Fifth HCT Information Technology Trends (ITT)*, pages 358–363, 2018.

[224] Jean Burgess. The iphone moment, the apple brand, and the creative consumer: From "hackability and usability" to cultural generativity. In *Studying mobile media*, pages 28–42. Routledge, 2012.

[225] D G Morin, Ester Gonzalez-Sosa, Pablo Perez, and Alvaro Villegas. Full body video-based self-avatars for mixed reality: from e2e system to user study. *Virtual Reality*, 27(3):2129–2147, 2023.

[226] David Cortes, Belen Bermejo, and Carlos Juiz. The use of cnns in vr/ar/mr/xr: a systematic literature review. *Virtual Reality*, 28(3):154, 2024.

[227] Polona Caserman, Augusto Garcia-Agundez, Robert Konrad, Stefan Göbel, and Ralf Steinmetz. Real-time body tracking in virtual reality using a vive tracker. *Virtual Reality*, 23:155–168, 2019.

[228] Dayoung Jeong, Seungwon Paik, YoungTae Noh, and Kyungsik Han. Mac: multi-modal, attention-based cybersickness prediction modeling in virtual reality. *Virtual Reality*, 27(3):2315–2330, 2023.

[229] Polona Caserman, Augusto Garcia-Agundez, and Stefan Göbel. A survey of full-body motion reconstruction in immersive virtual reality applications. *IEEE transactions on visualization and computer graphics*, 26(10):3089–3108, 2019.

[230] Oscar R Toasa, Yadira Semblantes, David Martínez, Paúl Baldeón, and Renato M Toasa. Virtual reality in e-commerce: Brief review of current state. In *International Conference on Marketing and Technologies*, pages 647–655. Springer, 2024.

[231] Aysu Erensoy, Anuradha Mathrani, Alexander Schnack, Jonathan Elms, and Nilufar Baghaei. Consumer behavior in immersive virtual reality retail environments: A systematic literature review using the stimuli-organisms-responses (s-o-r) model. *Journal of Consumer Behaviour*, 23(6):2781–2811, 2024.

[232] Shugang Li, Boyi Zhu, and Zhaoxu Yu. The impact of cue-interaction stimulation on impulse buying intention on virtual reality tourism e-commerce platforms. *Journal of Travel Research*, 63(5):1256–1279, 2024.

[233] Dipayan Biswas, Annika Abell, and Roger Chacko. Curvy digital marketing designs: virtual elements with rounded shapes enhance online click-through rates. *Journal of Consumer Research*, 51(3):552–570, 2024.

[234] Alexander Schnack, Malcolm J Wright, and Judith L Holdershaw. Does the locomotion technique matter in an immersive virtual store environment?–comparing motion-tracked walking and instant teleportation. *Journal of Retailing and Consumer Services*, 58:102266, 2021.

[235] Sharmeen M Saleem Abdullah and Adnan Mohsin Abdulazeez. Facial expression recognition based on deep learning convolution neural network: A review. *Journal of Soft Computing and Data Mining*, 2(1):53–65, 2021.

[236] Shekhar Singh. *Facial Expression Recognition Using Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) for Data Augmentation and Image Generation*. PhD thesis, University of Nevada, Las Vegas, 2023.

[237] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023.

[238] Stewart I Donaldson and Elisa J Grant-Vallone. Understanding self-report bias in organizational behavior research. *Journal of business and Psychology*, 17:245–260, 2002.

[239] Xiaozhi Yang and Ian Krajbich. Webcam-based online eye-tracking for behavioralresearch. *Judgment and Decision making*, 16(6):1485–1505, 2021.

[240] John L Andreassi. *Psychophysiology: Human behavior and physiological response*. Psychology press, 2010.

[241] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[242] Ho-Seung Cha and Chang-Hwan Im. Performance enhancement of facial electromyogram-based facial-expression recognition for social virtual reality applications using linear discriminant analysis adaptation. *Virtual Reality*, 26(1):385–398, 2022.

[243] Mohammed Aly, Abdullatif Ghallab, and Islam S Fathi. Enhancing facial expression recognition system in online learning context using efficient deep learning model. *IEEE Access*, 11:121419–121433, 2023.

[244] Bita Houshmand and Naimul Mefraz Khan. Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 70–75. IEEE, 2020.

[245] Ho-Seung Cha and Chang-Hwan Im. Improvement of robustness against electrode shift for facial electromyogram-based facial expression recognition using domain adaptation in vr-based metaverse applications. *Virtual Reality*, 27(3):1685–1696, 2023.

[246] Patrizia Cherubino, Ana C Martinez-Levy, Myriam Caratù, Giulia Cartocci, Gianluca Di Flumeri, Enrica Modica, Dario Rossi, Marco Mancini, and Arianna Trettel. Consumer behaviour through the eyes of neurophysiological measures: State-of-the-art and future trends. *Computational intelligence and neuroscience*, 2019(1):1976847, 2019.

[247] E E Miller and L N Boyle. Variations in road conditions on driver stress: insights from an on-road study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 864–1868, San Diego: Sept 2013, 2013.

[248] Erika E Miller. Effects of roadway on driver stress: An on-road study using physiological measures. Master's thesis, University of Washington, Seattle, WA, 2013. Available at http://hdl.handle.net/1773/23592.

[249] E E Miller and L N Boyle. Driver adaptation to lane keeping assistance systems: Do drivers become less vigilant? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 1934–1938, 2017.

[250] E E Miller and L N Boyle. Adaptations in attention allocation: Implications for takeover in an automated vehicle. *Transportation research part F: traffic psychology and behaviour*, 66:101–110, 2019.

[251] Dario Di Dario, Viviana Pentangelo, Maria Immacolata Colella, Fabio Palomba, and Carmine Gravino. Collecting and implementing ethical guidelines for emotion recogni-

tion in an educational metaverse. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 549–554, 2024.

[252] Sana Ullah, Jie Ou, Yuanlun Xie, and Wenhong Tian. Facial expression recognition (fer) survey: a vision, architectural elements, and future directions. *PeerJ Computer Science*, 10:e2024, 2024.