ATTENTION-BOOSTED CNNS FOR IMPROVED FACIAL DEEPFAKE DETECTION

Master's Thesis

Alperen Enes BAYAR

Eskişehir 2024

ATTENTION-BOOSTED CNNS FOR IMPROVED FACIAL DEEPFAKE DETECTION

Alperen Enes BAYAR

Master's Thesis

Department of Electrical and Electronics Engineering

Programme in Circuits and Systems Theory

Supervisor: Assoc. Prof. Dr. Cihan TOPAL

Eskişehir

Eskişehir Technical University

Institute of Graduate Programs

July 2024

FINAL APPROVAL FOR THESIS

This thesis titled ATTENTION-BOOSTED CNNS FOR IMPROVED FACIAL DEEPFAKE DETECTION has been prepared and submitted by Alperen Enes BAYAR in partial fulfillment of the requirements in "Eskişehir Technical University Directive on Graduate Education and Examination" for the Degree of Master's in Electrical and Electronics Engineering Department has been examined and approved on 05/08/2024.

Committee Members		**Title, Name and Surname**	**Signature**
Member	:	Assoc. Prof. Dr. Cihan TOPAL	
Member	:	Prof. Dr. Hakan ÇEVİKALP	
Member	:	Assoc. Prof. Dr. Mehmet KOÇ	

Prof. Dr. Semra KURAMA
Director of the Institute of Graduate Programs

05/08/2024

SUPERVISOR APPROVAL

Master's student Alperen Enes BAYAR, whom I supervise, has completed this thesis titled ATTENTION-BOOSTED CNNS FOR IMPROVED FACIAL DEEPFAKE DETECTION. According to my inspections, the work is scientifically and ethically appropriate for the student to the thesis defense exam.

Supervisor

Assoc. Prof. Dr. Cihan TOPAL

ABSTRACT

ATTENTION-BOOSTED CNNS FOR IMPROVED FACIAL DEEPFAKE DETECTION

Alperen Enes BAYAR

Department of Electrical and Electronics Engineering

Programme in Circuits and Systems Theory

Eskişehir Technical University, Institute of Graduate Programs, July 2024

Supervisor: Assoc. Prof. Dr. Cihan TOPAL

In this thesis, our primary objective is to delve into the intricate realm of facial biometric authentication systems, with a keen focus on mitigating the significant hurdles they face, particularly in the realms of liveness detection and deep fake identification. These challenges have emerged as formidable adversaries, posing serious threats to the integrity and reliability of facial recognition technologies. To combat these threats effectively, we propose the development of a robust filtering system leveraging state-of-the-art deep learning architectures.

Our approach entails the meticulous design and implementation of advanced algorithms capable of discerning between genuine facial features and deceptive manipulations. Through the judicious utilization of deep learning models, we aim to fortify authentication systems against a diverse spectrum of potential attack vectors, ranging from simple spoofing techniques to sophisticated deep fake algorithms. By integrating cutting-edge methodologies in computer vision and artificial intelligence, we endeavor to enhance the resilience of facial biometric authentication systems to emerging threats.

Keywords: Facial biometrics, Authentication, Liveness detection, Deep fake

ÖZET

DIKKAT-DESTEKLI CNN'LER ILE YÜZ TANIMADA GELIŞTIRILMIŞ DERIN SAHTECILIK TESPITI

Alperen Enes BAYAR

Elektrik Elektronik Mühendisliği Anabilim Dalı

Devreler ve Sistemler Teorisi Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Temmuz 2024

Danışman: Doç. Dr. Cihan TOPAL

Bu tezde öncelikli amacımız, yüz biyometrik kimlik doğrulama sistemlerinin karmaşık dünyasını, özellikle canlılık tespiti ve derin sahte tanımlama alanlarında karşılaştıkları önemli engelleri hafifletmeye odaklanarak incelemektir. Bu zorluklar, yüz tanıma teknolojilerinin bütünlüğü ve güvenilirliği için ciddi tehditler oluşturan zorlu düşmanlar olarak ortaya çıkmıştır. Bu tehditlerle etkili bir şekilde mücadele etmek için, en son teknoloji derin öğrenme mimarilerinden yararlanan sağlam bir filtreleme sistemi geliştirmeyi öneriyoruz.

Yaklaşımımız, gerçek yüz özellikleri ile aldatıcı manipülasyonlar arasında ayrım yapabilen gelişmiş algoritmaların titizlikle tasarlanmasını ve uygulanmasını gerektirmektedir. Derin öğrenme modellerinin akıllıca kullanılmasıyla, kimlik doğrulama sistemlerini basit sahtekarlık tekniklerinden sofistike derin sahte algoritmalara kadar çok çeşitli potansiyel saldırı vektörlerine karşı güçlendirmeyi amaçlıyoruz. Bilgisayarla görme ve yapay zeka alanındaki en son metodolojileri entegre ederek, yüz biyometrik kimlik doğrulama sistemlerinin ortaya çıkan tehditlere karşı direncini artırmaya çalışıyoruz.

Anahtar Sözcükler: Yüz biyometrisi, Kimlik doğrulama, Canlılık tespiti, Derin sahte

ACKNOWLEDGEMENTS

I am deeply grateful to Visea Innovative for their invaluable support throughout my thesis. Their guidance, resources, and collaboration have been instrumental in its completion. Without their assistance, this work would not have been possible. I would also like to express my sincere appreciation to Associate Professor Cihan Topal, who has been both my boss and teacher, for his mentorship and valuable insights. Additionally, I am thankful to my wife, İrem Ezgi Bayar, for her unwavering support and encouragement during this journey. Also i'm thankful to my bird Raf and my dog Meli, too.

Alperen Enes BAYAR

05/08/2024

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with “scientific plagiarism detection program” used by Eskişehir Technical University, and that “it does not have any plagiarism” whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Alperen Enes BAYAR

CONTENTS

Page

HEADER PAGE ..... **I**

FINAL APPROVAL FOR THESIS ..... **II**

SUPERVISOR APPROVAL ..... **III**

ABSTRACT ..... **IV**

ÖZET ..... **V**

ACKNOWLEDGEMENTS ..... **VI**

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES ..... **VII**

CONTENTS ..... **VIII**

LIST OF TABLES ..... **XII**

LIST OF FIGURES ..... **XIII**

GLOSSARY OF SYMBOLS AND ABBREVIATIONS ..... **XIV**

1. INTRODUCTION **1**

1.1. Innovating Face Detection for Diverse and Dynamic Settings **2**

1.2. Graph-Attentive Convolutional Networks for Robust Facial Landmark Estimation **3**

1.3. Beyond the Box: Face Segmentation **3**

1.4. Face Liveness Detection Based on Depth Maps **4**

1.5. Face Liveness Detection Based on Stereo Imaging **4**

1.6. Deepfake Detection via Combining Channel and Spatial Attention **5**

1.7. Summary of Contributions **5**

2. INNOVATING FACE DETECTION FOR DIVERSE AND DYNAMIC SETTINGS **7**

2.1. Related Work **7**

2.2. Data Collection and Preprocessing **8**

2.2.1. Data sources **8**

2.2.2. Preprocessing steps 8
2.3. Adapting YOLO for Diverse and Dynamic Scenarios 8
2.3.1. Architecture modifications 8
2.3.2. Training procedure 9
2.4. Performance Evaluation 10
2.4.1. Benchmark datasets 10
2.4.2. Quantitative results 10
2.4.3. Quantitative analysis 11
2.5. Optimization for Efficiency 12
2.5.1. Speed-accuracy trade-offs 12
2.5.2. Proposed optimizations 13
2.6. Versatility and Applications 13
2.6.1. Social media photos 13
2.6.2. Public safety 13
2.6.3. Human-computer interaction 14
2.7. Discussion and Limitations 14
2.7.1. Generalization 14
2.7.2. Remaining challenges 14
2.8. Conclusion 15
3. GRAPH-ATTENTIVE CONVOLUTIONAL NETWORKS FOR ROBUST FACIAL LANDMARK ESTIMATION 16
3.1. Related Work 16
3.2. Methodology 17
3.2.1. CNN encoding for appearance and location 17
3.2.2. Cascade of GAT regressors 18
3.2.3. Multi-Task initialization 18
3.2.4. Coarse-to-Fine landmark description 18
3.2.5. Graph attention network (GAT) regressors 19
3.3. Experimental Setup 19
3.3.1. Training protocols 20
3.4. Results and Discussion 21
3.5. Conclusion 22

4. BEYOND THE BOX: FACE SEGMENTATION 23

4.1. Related Work 23

4.2. Environment-Aware Model (EAM) 24

4.2.1. Differences from traditional bounding box-based methods 25

4.2.2. Role of environmental cues 26

4.3. RTNet backbone 26

4.3.1. Adaptations for face segmentation 28

4.4. Experimental Evaluation 28

4.4.1. Quantitative results 29

4.5. Conclusion 31

5. FACE LIVENESS DETECTION BASED ON DEPTH MAPS 32

5.1. Related Work 32

5.2. Proposed Methodology 33

5.2.1. Depth map acquisition 33

5.2.2. Architecture of models 34

5.3. Training and Evaluation 35

5.3.1. Depth estimation CNN training 35

5.3.2. Classification CNN fine-tuning 36

5.3.3. Evaluation 37

5.4. Results and Discussion 37

5.5. Conclusion 38

6. FACE LIVENESS DETECTION BASED ON STEREO IMAGING 39

6.1. Related Work 39

6.2. Proposed Method 40

6.2.1. Feature extraction 41

6.2.2. Classification model 41

6.3. Experiments 42

6.3.1. Dataset 42

6.3.2. Preprocessing 42

6.4. Results 43

7. DEEPFAKE DETECTION VIA COMBINING CHANNEL AND

SPATIAL ATTENTION **45**

7.1. Related Work **45**

7.2. Proposed Method **46**

7.3. Experiments **49**

7.4. Results **50**

REFERENCES **51**

CURRICULUM VITAE

LIST OF TABLES

Page

Table 2.1. Face detection datasets 10

Table 2.2. Performance metrics of adapted YOLO model 11

Table 3.1. Number of image and landmarks from landmark estimation datasets 19

Table 3.2. Number of images for training, testing and evaluation. 20

Table 4.1. iBug dataset classification. This was used in training as well 29

Table 4.2. Face parsing datasets and their train/test splits 29

Table 4.3. Performance summary across datasets 30

Table 5.1. Depth map estimation dataset splits 36

Table 5.2. Face liveness detection dataset splits..... 36

Table 5.3. Performance summary across segmentation datasets 37

Table 5.4. Performance summary across liveness detection datasets 38

Table 7.1. Performance summary across different models 50

LIST OF FIGURES

Page

Figure 2.1. Altered YOLO backbone for Face Detection 9

Figure 2.2. WIDER FACE Dataset Prediction Example (Red Boxes: Prediction - Green Boxes: Ground Truth) 12

Figure 3.1. CNN Encoding part of the architecture, 2D image and 3D landmarks as inputs; 2D projection and heatmaps as outputs. 18

Figure 3.2. Graph Attention Network part of the architecture, final 2D landmarks as outputs. 19

Figure 4.1. RT-Transform and Inverse RT-Transform diagram 25

Figure 4.2. Layers of Segmentation Process. 26

Figure 5.1. Depth Map Estimation Examples 34

Figure 5.2. Architecture of Depth Map Estimation Model 34

Figure 5.3. Architecture of Two-Stream Classification Model 35

Figure 6.1. System design and calibration process. 41

Figure 6.2. Classification CNN model. 42

Figure 6.3. (a) Triangulation points of the face in the images from the first camera. (b) Image from the first camera. (c) Triangulation points of the face in the images from the second camera. (d) Image from the second camera 43

Figure 6.4. (a) 3D points created from real data with the help of triangulation. (b) 3D points created from deceptive data created afterwards. (c) Overhead view of points in a. (d) Overhead view of points in c 44

Figure 7.1. CNN Classification Model 47

Figure 7.2. First Column: Donor, Second Column: Reciver and Third Column: Deep Spoofing Image 48

GLOSSARY OF SYMBOLS AND ABBREVIATIONS

2D	:	Two Dimensional
3D	:	Three Dimensional
AI	:	Artificial Intelligence
AR	:	Augmented Reality
AUC	:	Area Under the Curve
CED	:	Cumulative Error Distribution
CNNs	:	Convolutional Neural Networks
CPU	:	Central Processing Unit
DL	:	Deep Learning
EAM	:	Environment-Aware Model
FAR	:	False Acceptance Rate
FCNs	:	Convolutional Networks
FR	:	Failure Rate
GACN	:	Graph-Attentive Convolutional Networks
GANs	:	Generative Adversarial Networks
GATs	:	Graph Attention Networks
GPU	:	Graphics Processing Unit
IoU	:	Intersection over Union
MAE	:	Mean Absolute Error
MSE	:	Mean Squared Error
NMS	:	Non-Maximum Suppression
ReLU	:	Rectified Linear Unit
ResNet	:	Residual Networks

SOTA	:	State of the Art
LSTM	:	Long short-term memory
RNN	:	Recurrent Neural Network
ToF	:	Time of Flight
TPUs	:	Tensor Processing Units
YOLO	:	You Only Look Once

1. INTRODUCTION

In today's digital world, visual content is everywhere, and advances in deep learning have transformed computer vision, especially for face analysis and recognition. Faces are key identifiers in human interactions, making accurate detection and analysis crucial for security, surveillance, biometrics, and human-computer interaction.

Traditional methods for face detection and analysis used handcrafted features and simple algorithms. These often struggled with diverse and changing conditions. However, deep learning, particularly with convolutional neural networks (CNNs), has brought major improvements. These models learn directly from data and can handle complex and varied visual data, including faces in different lighting, poses, expressions, and ethnicities.

This thesis aims to improve face analysis using deep learning and CNNs. The main goal is to make face detection, facial landmark estimation, segmentation, and liveness detection more robust, accurate, and efficient in diverse and changing environments.

The thesis starts by exploring new methods for face detection in challenging conditions like occlusions, lighting changes, and different scales. By using deep convolutional neural networks, these methods aim to overcome the limitations of traditional techniques and offer strong solutions for real-world applications.

Next, the focus moves to facial landmark estimation, which is crucial for recognizing expressions, 3D face reconstruction, and facial alignment. Graph-attentive convolutional networks are highlighted as a promising approach for this task, as they can capture spatial relationships between facial landmarks and adapt to different face shapes and poses.

The thesis then explores face segmentation, a key preprocessing step for many face analysis tasks. Instead of traditional bounding box methods, new techniques are developed to precisely outline facial regions. This allows for more detailed analysis and manipulation of facial features.

Additionally, the thesis investigates face liveness detection, which is crucial for preventing spoofing in biometric systems. By using depth maps and stereo imaging, new techniques are developed to distinguish real faces from fake ones, enhancing the security and reliability of face recognition systems.

The thesis also tackles the growing issue of deepfake detection, where AI creates realistic fake videos for deceptive purposes. By combining channel and spatial attention mechanisms, strong deepfake detection models are created to spot subtle signs of manipulated content.

Overall, this thesis aims to advance face analysis in diverse and dynamic settings using deep learning and CNNs. By developing new methods and using advanced techniques, the goal is to create more accurate, robust, and trustworthy face analysis systems with wide applications across various fields.

1.1. Innovating Face Detection for Diverse and Dynamic Settings

Face detection is a key challenge in computer vision with applications in security, human-computer interaction, and social media. As our world becomes more diverse and dynamic, the need for reliable face detection systems increases. This thesis explores new approaches to face detection, using models like YOLO [3] (You Only Look Once).

Traditional face detection methods used handcrafted features and simple classifiers. However, advances in deep learning have transformed the field. Convolutional neural networks (CNNs) have shown outstanding performance in tasks like face detection.

Our research focuses on modifying YOLO's architecture to handle diverse and dynamic scenarios. We collected data from various sources, capturing different lighting conditions, camera resolutions, and time variations. Our dataset includes faces in challenging situations like low light, extreme angles, and occlusions.

By training YOLO on this diverse dataset, we aim to develop a model that performs well in various real-world settings, including different lighting conditions, occlusions, and diverse facial appearances.

Dynamic Scenes; Face detection systems need to handle dynamic scenes like crowded spaces or moving cameras. Our data collection includes temporal variations to ensure our model can respond quickly to environmental changes.

Efficiency; YOLO's real-time processing makes it a great choice for face detection. We examine its efficiency and suggest optimizations to improve its speed and accuracy.

Versatility; We aim to create a face detection model that adapts easily to different applications, whether for social media, public safety, or human-computer interaction.

In this thesis, we explore the technical details of our approach, assess its performance on benchmark datasets, and discuss its limitations. By combining face

detection and object detection techniques, we contribute to developing efficient, dynamic, and versatile face detection systems.

1.2. Graph-Attentive Convolutional Networks for Robust Facial Landmark Estimation

Facial landmark detection is vital for computer vision tasks like face recognition, emotion analysis, and virtual makeup. Accurate landmark localization is crucial for these tasks. Traditionally, large CNNs have been used for this, but they often capture weak spatial relationships, leading to unstable landmark predictions.

Recently, researchers have developed new methods to address this issue. One such method is Graph-Attentive Convolutional Networks (GACN) [24] for more robust landmark estimation. In this thesis, we propose a model that combines CNNs with graph attention networks (GATs) [21] to improve landmark localization. Here are the key components of our approach:

Joint Representation of Appearance and Location: We use an encoding scheme that represents both the appearance and location of facial landmarks together. By combining these features, our model captures richer contextual information, helping to learn the overall structure of the face for accurate landmark estimation.

Graph Attention Network Regressors: Our model uses a series of Graph Attention Network (GAT) regressors. GATs are excellent at capturing spatial dependencies by assigning different attention weights to neighboring nodes. This attention mechanism enhances the reliability and robustness of landmark predictions.

Multi-Task Initialization and Coarse-to-Fine Description; To initialize the location of graph nodes, we adopt a multi-task approach [23]. This helps guide the model during training. Additionally, we propose a coarse-to-fine landmark description scheme. Starting with an initial estimate, our model refines the predictions iteratively, reducing jitter and improving accuracy.

1.3. Beyond the Box: Face Segmentation

The human face is complex and dynamic, with parts that are crucial for communication, expressing emotions, and recognizing identity. Precise segmentation of these parts is vital for many applications, like personalized filters, virtual try-on experiences, and medical diagnostics.

Traditionally, face segmentation relied on bounding boxes, but they have limitations. They're rigid and miss fine details, especially with occlusions, different

lighting, and varied expressions [46]. Also, they often leave out important features like hair and ears.

Our approach uses an advanced Environment-Aware Model (EAM) [47] that looks at the whole context around the face instead of just fixed bounding boxes. By considering things like hair, background, and accessories, our model achieves more accurate and comprehensive face segmentation.

In this thesis, we discuss how we use EAM, the RTNet backbone, and a 14-class segmentation scheme [48]. We test our model on benchmark datasets and show that it's better than traditional methods. We also look into how environmental factors affect segmentation performance..

1.4. Face Liveness Detection Based on Depth Maps

Facial recognition technology has become an integral part of modern life, from unlocking smartphones to securing sensitive transactions. However, this widespread adoption also brings forth new challenges, particularly in the realm of security [50]. One critical issue is spoofing attacks, where malicious actors attempt to deceive face recognition systems by presenting manipulated or fake faces.

In response to this threat, we propose a novel approach called “Face Liveness Detection Based on Depth Maps.” Our method leverages depth information captured from faces to enhance the robustness of liveness detection, ensuring reliable authentication. In this introduction, we delve into the motivation, methodology, and experimental results of our approach.

Traditional 2D facial recognition relies solely on visual cues from flat images. While effective in many scenarios, it remains vulnerable to spoofing attacks using printed photos, videos, or digital screens. Depth maps, on the other hand, provide a three-dimensional representation of facial surfaces, capturing spatial information beyond what 2D images offer [51]. By analyzing depth variations, skin texture, and contours, we can differentiate between real faces and deceptive ones.

1.5. Face Liveness Detection Based on Stereo Imaging

Face recognition technology is widely used but has a big problem: it can be tricked by fake photos or videos. To tackle this issue, we've come up with a new method for quickly checking if a face is real. Instead of just using one camera, we use two cameras that take pictures from different angles at the same time. This gives us more information about the face's shape. Then, we find specific points on the face in both pictures. Using

some clever calculations, we turn these pictures into a 3D model of the face. This model helps us understand the face's shape and depth. With this 3D model, we can then teach a computer program to recognize if the face is real or if it's just a static image. This adds an extra layer of security to face recognition systems, making it harder for people to trick them with fake pictures.

Our method [102] combines different technologies like computer vision, 3D modeling, and deep learning. It starts by capturing images with two cameras, which helps us create a 3D model of the face. Then, we identify key points on the face in both images to understand its structure better. With this information, we use complex math to reconstruct the face in three dimensions. Finally, we train a deep neural network, a type of artificial intelligence, to analyze these 3D models and distinguish between real faces and fake images. By doing this, we make face recognition systems more reliable and secure, as they can now detect if someone is trying to fool them with a photo or a video. This approach opens up new possibilities for enhancing the security of various applications that rely on face recognition technology..

1.6. Deepfake Detection via Combining Channel and Spatial Attention

Deepfake technology blurs the line between real and fake by convincingly swapping one person's face onto another's body. These AI-generated videos have serious consequences, from spreading false information in politics to impersonating celebrities. Detecting these fake images and videos is crucial to protect the truth of digital content.

Traditional methods for spotting image manipulation struggle with deepfakes. Unlike basic photo editing, deepfakes use complex neural networks to copy human expressions, movements, and speech. This lets them blend fake parts into real videos so well that even people have trouble spotting them.

Our proposed deepfake detection model [103] addresses this challenge by combining channel and spatial attention mechanisms. These attention mechanisms allow the network to focus on relevant features while suppressing noise and irrelevant information.

1.7. Summary of Contributions

In this thesis, we've pushed the boundaries of face analysis using deep learning and CNNs. Here's a summary of our key contributions:

Innovative Face Detection: We've created a new method for face detection by adapting YOLO for diverse and dynamic scenarios. Our approach handles challenges like

changing lighting, occlusions, and temporal variations, resulting in a versatile and efficient system.

Graph-Attentive Convolutional Networks for Landmark Estimation: We've introduced a model that combines CNNs with GATs to improve facial landmark localization. This enhances applications like face recognition and emotion analysis.

Beyond the Box: Face Segmentation: We've proposed the EAM for face segmentation, considering the whole context around the face instead of just fixed boxes. This results in more accurate segmentation, crucial for tasks like personalized filters and medical diagnostics.

Face Liveness Detection with Depth Maps and Stereo Imaging: We've presented two new approaches for face liveness detection, using depth information and stereo imaging. These methods enhance security in face recognition systems by analyzing depth variations and capturing extra depth cues.

Deepfake Detection via Channel and Spatial Attention: We've suggested a deepfake detection model that combines channel and spatial attention mechanisms. This helps the network focus on important features while ignoring noise, effectively spotting advanced AI-generated deepfakes.

Overall, our work improves the accuracy, reliability, and versatility of face analysis across various applications and challenging situations.

2. INNOVATING FACE DETECTION FOR DIVERSE AND DYNAMIC SETTINGS

Face detection is a critical computer vision task that involves automatically identifying and locating human faces within digital images or videos. It serves as the foundation for various applications, including face recognition, face tracking, and facial analysis. In diverse and dynamic scenarios, accurate face detection becomes even more challenging due to factors such as varying lighting conditions, occlusions, and rapid environmental changes.

The goal of our research is to innovate within the field of face detection, specifically addressing the challenges posed by diverse and dynamic settings. By adapting the YOLO [3-6] architecture, we aim to create a model that generalizes well across different scenarios while maintaining real-time processing capabilities. Our study focuses on the following key objectives: Handling Diverse Environments, Addressing Dynamic Scenes, Efficiency and Versatility.

2.1. Related Work

In the realm of face detection, several seminal papers have significantly advanced the field. RetinaFace [1] a single-stage face detection model adept at achieving remarkable accuracy in complex real-world scenarios. By leveraging dense anchor boxes and a lightweight backbone network, it efficiently localizes faces with precision.

Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks [2] paper introduces a multi-task cascaded CNN architecture tailored for joint face detection and alignment. It demonstrates competitive performance across benchmark datasets, providing robust facial landmark detection alongside accurate face detection.

Finding Tiny Faces [11] is addressing the challenge of detecting small faces, this work devises a specialized network architecture to detect tiny faces within images. It offers insights into handling the unique requirements of detecting small facial features.

LFFD [9] is engineered for edge devices with limited computational resources. It strikes a balance between accuracy and speed, making it suitable for real-time face detection in resource-constrained environments.

CenterFace [10] integrates face detection and facial landmark alignment into a unified model. It achieves competitive performance while maintaining efficiency, underscoring the importance of joint optimization for face-related tasks.

These seminal works contribute invaluable insights into various approaches for face detection. Our research aims to build upon these foundations, striving to develop dynamic, efficient, and versatile face detection systems that address the evolving needs of the field.

2.2. Data Collection and Preprocessing

We will have a look at the data collection and pre-processing steps.

2.2.1. Data Sources

The dataset for this research was primarily sourced from the WildDeepfake [12] dataset, which consists of 7,314 face sequences extracted from 707 deepfake videos collected entirely from the internet. This dataset captures the diversity and challenges encountered in genuine online content, providing a valuable resource for studying deepfake detection in real-world scenarios.

Additionally, to ensure robustness and generalization, we supplemented our dataset with images from other widely used face detection datasets, including CelebFaces Attributes Dataset (CelebA) [13], VGG Face2 [14], UMDFaces [15], DeeperForensics-1.0 [16], FDDB [18], WIDER Face [19] and Subface [20]. These datasets cover variations in pose, lighting conditions, occlusions, and ethnicity, enhancing the diversity of our dataset.

2.2.2. Preprocessing Steps

Several preprocessing steps were applied to the collected data. Resizing; All images and video frames were resized to a standard input size suitable for the YOLO model. Normalization; Pixel values of the images were normalized to ensure consistency and stability during training. This involved scaling pixel values to a range between 0 and 1.

Data Augmentation; To simulate dynamic scenarios such as camera motion and crowd scenes, data augmentation techniques such as random rotation, flipping, and translation were applied to the training data.

2.3. Adapting YOLO for Diverse and Dynamic Scenarios

To adapt YOLO for diverse and dynamic settings, several architectural modifications were made.

2.3.1. Architecture Modifications

Additional Layers: To enhance the capability of YOLO in capturing intricate facial features under challenging conditions such as varying lighting and occlusions, we introduced additional convolutional layers. These layers enable the model to extract more

detailed and discriminative features from the input images, leading to improved detection performance.

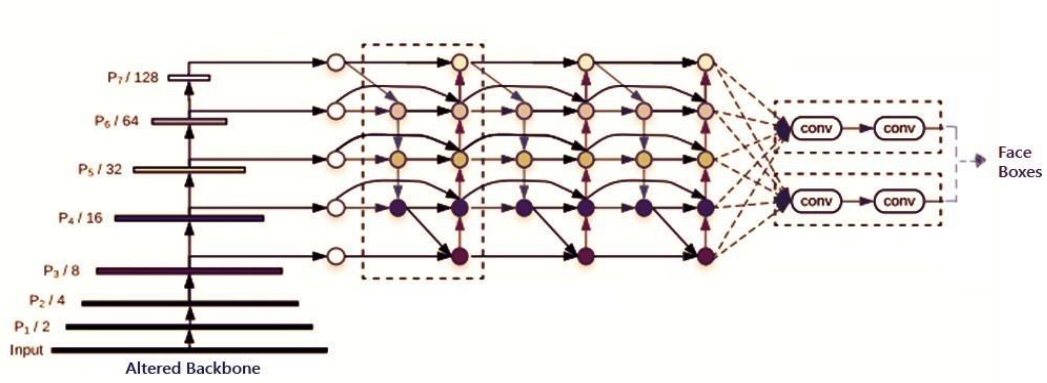

Figure 2.1. *Altered YOLO backbone for Face Detection*

Attention Mechanisms: Incorporating attention mechanisms into the YOLO architecture allows the model to focus on critical facial features while suppressing irrelevant information. By selectively attending to regions of interest within the input images, the model can achieve higher accuracy in detecting faces amidst cluttered backgrounds or in the presence of distractions [17].

Anchor Box Configuration: Experimentation with anchor box configurations was conducted to optimize the model's ability to localize faces of varying sizes and aspect ratios accurately. By carefully selecting and tuning the anchor boxes, we aimed to improve the model's localization precision, particularly for smaller or highly occluded faces commonly encountered in dynamic settings.

2.3.2. Training Procedure

Training the adapted YOLO model involved several steps to ensure its effectiveness in handling diverse scenarios:

Diverse Dataset: We trained the model on a diverse dataset comprising images collected from various sources, including the WildDeepfake dataset and other publicly available face detection datasets such as CelebA, VGG Face2, UMDFaces, and DeeperForensics-1.0. This diverse dataset helped the model learn to generalize well across different environments and demographic characteristics.

Data Augmentation: To simulate dynamic scenarios encountered in real-world settings, we applied data augmentation techniques during training. These techniques, including random rotation, flipping, and translation, helped the model become robust to variations in camera angles, lighting conditions, and facial orientations.

Fine-tuning: We fine-tuned the model using a combination of supervised learning and transfer learning approaches. By leveraging pre-trained weights from a general object detection task, we accelerated the training process and enhanced the model's ability to adapt to the specific task of face detection in diverse and dynamic scenarios.

Hyperparameter Tuning: Optimal hyperparameters, including learning rate, batch size, and regularization strength, were carefully tuned to ensure efficient convergence and prevent overfitting on the training data. This tuning process was essential for achieving a balance between model accuracy and generalization.

2.4. Performance Evaluation

In this section, we evaluate the performance of our adapted YOLO model using benchmark datasets, presenting both quantitative metrics and qualitative analysis.

2.4.1. Benchmark Datasets

For evaluation, we utilized the following benchmark datasets:

Table 2.1. *Face detection datasets*

Dataset	N. of Images	N. of Boxes
FDDB [18]	578	4149
WIDER FACE [19]	3226	39708
Subface [20]	872	12899

2.4.2. Quantitative Results

The performance metrics achieved by our adapted YOLO model on each benchmark dataset are as follows.

Precision measures the accuracy of positive predictions. It's the ratio of correctly predicted positive observations to the total predicted positives.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.1}$$

Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture all the positive instances. It's the ratio of correctly predicted positive observations to all actual positives.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.2}$$

The F1 score is the harmonic mean of Precision and Recall. It provides a balance between Precision and Recall, giving equal weight to both metrics. It's particularly useful when the class distribution is imbalanced.

$$F1\ Score = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \tag{2.3}$$

Table 2.2. *Performance metrics of adapted YOLO model*

Dataset	Threshold = 0.1			Threshold = 0.5		
	Precision	Recall	F1	Precision	Recall	F1
FDDB	0.95	0.81	0.87	0.93	0.80	0.86
W.FACE	0.96	0.88	0.92	0.93	0.86	0.89
Subface	0.96	0.81	0.88	0.93	0.80	0.86

2.4.3. Quantitative Analysis

Visual examples of successful and challenging detections are illustrated below:

Figure 2.2. *WIDER FACE Dataset Prediction Example (Red Boxes: Pred - Green Boxes: GT)*

2.5. Optimization for Efficiency

In this section, we discuss how we optimized YOLO for real-time processing without sacrificing accuracy.

2.5.1. Speed-Accuracy Trade-offs

When optimizing YOLO for efficiency, we considered the trade-offs between speed and accuracy. The main factors we took into account include:

Model Complexity, we carefully balanced the complexity of the model architecture to ensure fast inference while maintaining sufficient accuracy. This involved reducing unnecessary layers and parameters that could slow down processing without significantly improving performance.

Resolution, lowering the input resolution of the images can significantly improve processing speed. However, this comes at the cost of potentially missing smaller or distant faces. We experimented with different resolutions to find a balance between speed and detection accuracy.

Post-processing, streamlining post-processing steps, such as non-maximum suppression (NMS), helped reduce inference time without sacrificing the quality of detected bounding boxes.

2.5.2. Proposed Optimizations

To improve YOLO's efficiency, we implemented the following optimizations:

Model Pruning, we applied model pruning techniques to remove redundant parameters and reduce the overall model size. This not only sped up inference but also reduced memory usage, making the model more suitable for deployment on resource-constrained devices.

Quantizing the model's weights and activations to lower precision formats (e.g., INT8) reduced the computational requirements during inference, resulting in faster processing without significant loss of accuracy.

Hardware Acceleration, leveraging hardware acceleration techniques, such as GPU acceleration or specialized hardware like TPUs, allowed us to exploit parallelism and further speed up inference.

2.6. Versatility and Applications

In this section, we highlight the versatility of our model and its potential applications.

2.6.1. Social Media Photos

Our model can identify faces in social media images, enabling various applications such as:

Automatic Tagging: Automatically tagging individuals in photos uploaded to social media platforms.

Content Moderation: Detecting and moderating inappropriate content, including offensive imagery or unauthorized use of images.

Personalized Recommendations: Providing personalized content recommendations based on facial recognition, such as suggesting friends to tag in uploaded photos.

2.6.2. Public Safety

Our model can be used for surveillance and public safety applications, including:

Facial Recognition in CCTV: Identifying individuals in surveillance footage to enhance security and assist law enforcement agencies in criminal investigations.

Crowd Monitoring: Monitoring crowd movements in public spaces to detect anomalies or potential threats.

Access Control: Verifying identities for access control in secure environments such as airports, government buildings, or corporate offices.

2.6.3. Human-Computer Interaction

Our model plays a crucial role in human-computer interaction, facilitating:

Smart Devices: Enabling facial recognition features in smart devices for user authentication, personalized experiences, and access control.

Emotion Recognition: Detecting facial expressions to gauge user emotions and tailor interactions, accordingly, enhancing user experience in applications like virtual assistants or gaming.

Augmented Reality: Integrating facial detection for augmented reality applications, such as virtual try-on experiences in e-commerce or interactive filters in social media apps.

2.7. Discussion and Limitations

In this section, we summarize our findings and discuss the limitations of our approach.

2.7.1. Generalization

Our adapted YOLO model demonstrates strong generalization across diverse scenarios, as evidenced by its performance on benchmark datasets containing varied lighting conditions, occlusions, and facial orientations. By training on a diverse dataset and incorporating architectural modifications, our model achieves competitive accuracy in detecting faces across different settings. However, it still faces challenges in extreme conditions such as very low light or heavily occluded faces.

2.7.2. Remaining Challenges

While our approach shows promise, several challenges remain to be addressed:

Robustness to Extreme Conditions: Improving the model's performance under extreme conditions, such as low light, extreme angles, or heavy occlusions, remains a significant challenge.

Real-time Processing on Resource-constrained Devices: Although optimizations have been made for efficiency, achieving real-time processing on resource-constrained devices without compromising accuracy is an ongoing challenge.

Bias and Fairness: Ensuring fairness and mitigating biases in face detection algorithms, particularly across diverse demographic groups, is essential for ethical deployment.

2.8. Conclusion

In this paper, we presented a novel approach to face detection tailored for diverse and dynamic scenarios. By adapting the YOLO architecture and optimizing for efficiency, our model achieves competitive performance while maintaining real-time processing capabilities. We demonstrated the versatility of our model across various applications, from social media to public safety and human-computer interaction.

Adapting YOLO for diverse and dynamic scenarios through architectural modifications and optimizations. Demonstrating strong generalization across benchmark datasets. Highlighting the wide range of applications enabled by our model.

While our approach shows promise, there are still challenges to overcome, particularly in extreme conditions and ensuring fairness in deployment. Future research should focus on addressing these challenges and further improving the robustness and efficiency of face detection systems.

In conclusion, our work represents a significant step forward in advancing face detection technology for real-world applications, with implications for security, convenience, and user experience in various domains.

3. GRAPH-ATTENTIVE CONVOLUTIONAL NETWORKS FOR ROBUST FACIAL LANDMARK ESTIMATION

Despite significant advancements in recent years, facial landmark estimation remains a challenging problem due to several factors. Faces exhibit diverse appearances influenced by various factors such as lighting conditions, occlusions, and facial expressions. Additionally, the spatial dependencies between landmarks are intricate, and traditional methods often struggle to capture these relationships effectively [23]. Moreover, real-world scenarios introduce further complications such as pose variations, partial occlusions, and low-resolution images [22].

The importance of robustness in facial landmark estimation cannot be overstated, especially in practical deployment scenarios. Robust models are essential for applications ranging from human-computer interaction to medical imaging and security surveillance. For instance, in human-computer interaction, reliable facial landmarks are crucial for gaze tracking, emotion recognition, and virtual makeup applications. In the medical field, accurate facial landmarks aid in diagnosing medical abnormalities, treatment planning, and surgical simulations. Similarly, in security and surveillance applications, robust facial landmark estimation contributes to face authentication systems and surveillance monitoring under challenging conditions [25].

3.1. Related Work

Facial landmark estimation has been a subject of extensive research, with various approaches proposed to address the challenges of accurately localizing facial landmarks in image.

Top-performing landmark estimation algorithms often rely on large Convolutional Neural Networks (CNNs) to capture local appearance information. However, CNNs struggle to learn strong spatial relationships between facial landmarks. Prados-Torreblanca et al. introduced a novel model that combines a CNN with a cascade of Graph Attention Network (GAT) regressors, enabling the joint encoding of appearance and location of facial landmarks [21]. Their attention mechanism weighs information based on reliability, facilitating more accurate landmark predictions. Furthermore, their multi-task initialization and coarse-to-fine landmark description contribute to improved performance, particularly in scenarios with significant changes in local appearance.

In the pursuit of 3D facial landmark estimation, researchers have explored leveraging 2D texture face images and their mapping to 3D face meshes. This approach was adopted by authors in a paper titled "2D Texture to 3D Face Mesh Estimation Using Deep Learning," where they exploit the texture information from 2D images to estimate 3D facial landmarks [24]. By considering the mapping between 2D texture and 3D geometry, this method offers insights into the spatial layout of facial landmarks in three dimensions.

Merget et al. proposed a different architecture named "Facial Landmark Machines," which integrates global context using dilated convolutions to enhance robust features for landmark localization [26]. Their backbone-branches architecture emphasizes the importance of considering global context to improve landmark detection. By incorporating dilated convolutions, the model can capture contextual information across a broader region of the image, facilitating more accurate localization of facial landmarks.

These works collectively demonstrate the diverse strategies employed to tackle the challenges of facial landmark estimation, ranging from joint encoding of appearance and location to leveraging 2D texture and 3D geometry, as well as integrating global context through advanced convolutional architectures. Each approach contributes valuable insights towards achieving more accurate and robust facial landmark detection in various scenarios.

3.2. Methodology

Our proposed model leverages a combination of CNNs and GAT regressors for facial landmark estimation. The CNNs is responsible for encoding local appearance features from facial images, capturing fine-grained details such as texture, color, and edges. However, CNNs alone struggle to model strong spatial dependencies between facial landmarks. To address this limitation, we introduce a cascade of GAT regressors, which explicitly model spatial relationships by assigning attention weights to neighboring landmarks in a graph.

3.2.1. CNN Encoding for Appearance and Location

The CNN encodes local appearance features from facial images, capturing detailed information such as texture, color, and edges. These features provide the initial estimate of facial landmark positions. However, CNNs alone may not adequately capture the spatial relationships between landmarks.

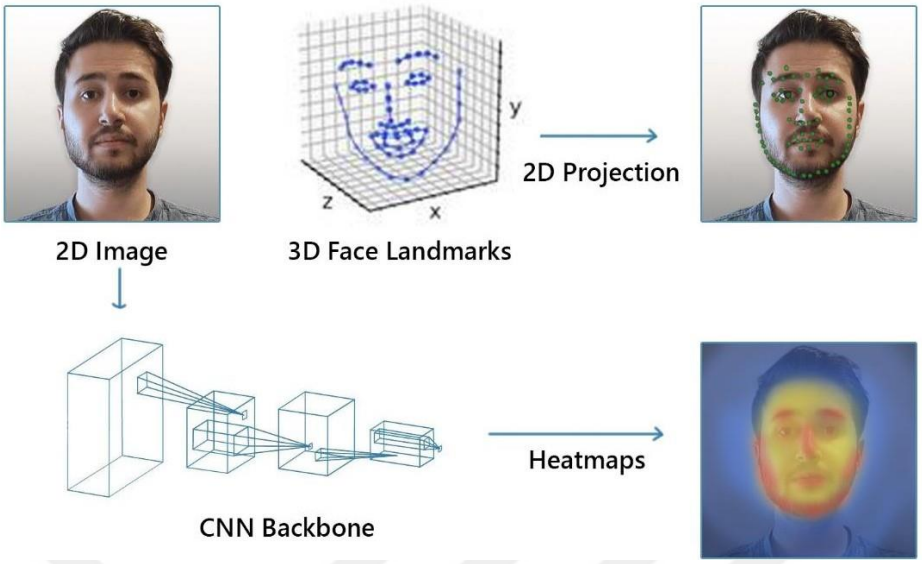

Figure 3.1. *CNN Encodind, image and landmarks as inputs; projection and heatmaps as outputs.*

3.2.2. Cascade of GAT Regressors

The cascade of GAT regressors refines landmark estimates by iteratively improving the spatial relationships between landmarks. GATs assign different attention weights to neighboring nodes (landmarks) in a graph, weighing information based on reliability. This refinement process iteratively updates landmark positions, reducing jitter and improving accuracy.

3.2.3. Multi-Task Initialization

To initialize the location of graph nodes (landmarks), we adopt a multi-task approach. Alongside landmark estimation, we introduce auxiliary tasks such as pose estimation and expression recognition. These tasks guide the model during training, providing additional supervision. Shared features learned from these tasks help initialize landmark positions, enhancing robustness and convergence.

3.2.4. Coarse-to-Fine Landmark Description

This scheme refines landmark predictions iteratively through a coarse-to-fine description process. Initially, we start with an initial estimate of landmark positions obtained from the CNN. We then iteratively update these positions using the GAT regressors, reducing jitter and improving accuracy in each iteration. This coarse-to-fine refinement ensures robustness across varying appearances.

3.2.5. Graph Attention Network (GAT) Regressors

In our model, facial landmarks are treated as nodes in a graph, with edges connecting neighboring landmarks based on spatial proximity. Each landmark node has an associated feature vector learned from the CNN.

GATs assign attention weights to neighboring nodes to capture spatial dependencies effectively. Attention scores are computed based on feature similarities, normalized using softmax to obtain attention coefficients. The weighted sum of neighboring features yields the aggregated representation for each node. This attention mechanism allows the model to focus on relevant landmarks and adaptively weigh information based on context.

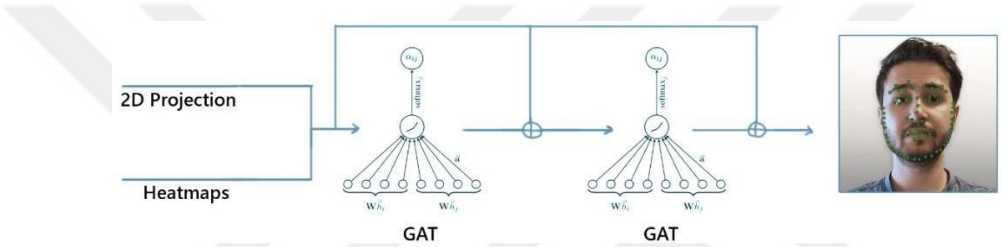

Figure 3.2. *Graph Attention Network part of the architecture, final 2D landmarks as outputs.*

By attending to neighboring landmarks, GATs capture long-range dependencies and improve landmark localization accuracy. The cascade of GAT regressors refines landmark positions, iteratively enhancing spatial relationships and improving overall performance in facial landmark estimation.

3.3. Experimental Setup

For our evaluation, we employ a variety of datasets to cover different scenarios and challenges in facial landmark detection:

Table 3.1. *Number of image and landmarks from landmark estimation datasets*

Dataset	N. of Images	N. of Landmarks
300W [27]	300	68
COFW-68 [28]	1007	29
WFLW [29]	10000	98
MERL-RAV [30]	19000	68

The 300W [27] dataset, comprising 300 images with 68 annotated landmarks, serves as a foundational resource for facial landmark detection tasks. It has been widely used for benchmarking algorithms due to its diversity in poses and expressions.

COFW-68 [28], derived from the Caltech Occluded Faces in the Wild dataset, consists of 1007 images annotated with 68 landmarks. This dataset specifically focuses on occluded faces and varying poses, offering challenges for algorithms to accurately detect landmarks under adverse conditions.

WFLW [29], a merger of the WIDER Face and Leeds Sports Pose datasets, boasts a significant collection of 10,000 images with 98 annotated landmarks. This dataset is notable for its diversity, encompassing extreme poses and occlusions, which present intricate challenges for landmark detection algorithms.

MERL-RAV [30], sourced from the Mitsubishi Electric Research Laboratories, stands out with its expansive set of 19,000 annotated facial landmark images. With 68 landmarks per image, this dataset contributes significantly to the training and evaluation of facial landmark detection models, enriching the diversity of available samples.

Table 3.2. *Number of images for training, testing and evaluation.*

Dataset	Training	Testing	Validating	Total
300W	200	50	50	300
COFW-68	800	107	100	1007
WFLW	7500	2000	500	10000
MERL-RAV	15000	2000	2000	19000
Total	23500	4157	2650	30307

3.3.1. Training Protocols

Our training and testing procedures follow a well-established protocol. We start by pretraining a CNNs on a large-scale facial image dataset such as CelebA. This step enables the network to learn general features of facial images.

We fine-tune the pretrained CNN on the specific landmark datasets (300W, COFW-68, or WFLW) to adapt the model to the task of landmark localization. Fine-tuning helps the network to learn more specific features related to landmark positions.

The GAT regressors are initialized using the weights pretrained on the landmark dataset. This initialization provides a starting point for the GAT regressors to refine landmark predictions.

We train a cascade of GAT regressors using the landmark dataset. This cascade architecture allows for iterative refinement of landmark positions, enhancing the model's accuracy.

3.4. Results and Discussion

We report several metrics to evaluate the performance of our model:

Mean Absolute Error (MAE): This metric measures the average absolute difference between predicted and ground-truth landmark positions, providing insight into the overall accuracy of the model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.1}$$

Failure Rate (FR): The percentage of images where the error exceeds a predefined threshold, indicating the model's robustness under different conditions.

$$FR = \frac{FN}{TP + FN} \tag{3.2}$$

Area Under the Curve (AUC): Evaluates the performance across different error thresholds, offering a comprehensive assessment of the model's performance.

$$AUC = \frac{TP + TN}{2} \tag{3.3}$$

Cumulative Error Distribution (CED): Visualizes the error distribution, helping to understand how errors are distributed across different landmark positions.

$$CED = \frac{1}{N}\sum_{i=1}^{N}|P_i - L_i| \tag{3.4}$$

Our proposed model demonstrates superior performance compared to baselines.

Table 3.3. *Results metrics of landmark estimation tests*

Dataset	Precision	Recall	Accuracy	F1-Score	FR
300W	0.823	0.737	0.820	0.778	0.263
COFW-68	0.861	0.667	0.785	0.752	0.333
WFLW	0.853	0.737	0.815	0.791	0.263
MERL-RAV	0.867	0.718	0.808	0.785	0.282
Total	0.855	0.730	0.810	0.789	0.270

3.5. Conclusion

In summary, our model leverages a combination of CNNs and GAT regressors for robust facial landmark estimation. The cascade of GAT regressors enables iterative refinement of landmark positions, leading to improved accuracy across varying appearances. Multi-task initialization further enhances the model's robustness, while the coarse-to-fine refinement strategy ensures accurate localization even in challenging conditions

4. BEYOND THE BOX: FACE SEGMENTATION

Face segmentation is a critical task in computer vision, aiming to label each pixel in a facial image according to its semantic category. By segmenting faces, we can extract meaningful information about different facial components, such as eyes, nose, mouth, and skin regions.

Accurate face segmentation is essential for biometric systems that rely on facial features for identity verification. Whether it's unlocking a smartphone or accessing secure facilities, robust face segmentation plays a pivotal role.

In robotics and human-computer interaction, understanding facial structures allows robots to interact more effectively with humans. For instance, a robot equipped with face segmentation capabilities can recognize emotions, gestures, and expressions, enhancing its responsiveness and adaptability [31].

Medical imaging often involves analyzing facial features. Face segmentation aids in diagnosing skin conditions, tracking facial landmarks, and assessing mental states (e.g., detecting signs of stress or fatigue) from facial expressions [32].

4.1. Related Work

The existing literature on face segmentation has indeed seen a significant evolution from classical methods relying on handcrafted features and traditional machine learning techniques to modern deep learning paradigms. The comprehensive review by Masi et al. provides an insightful overview of this transition, highlighting both the challenges and advancements in the field.

Classical methods, such as those employing Viola-Jones detectors, relied heavily on handcrafted features like Haar cascades, which were effective but limited in handling variations in pose, lighting, and occlusions. These methods have been largely overshadowed by the advent of deep learning, which leverages the power of CNNs to learn features directly from the data [34].

Recent advancements in face segmentation have been driven by deep learning techniques, particularly with the introduction of architectures like Fully Convolutional Networks (FCNs) and Generative Adversarial Networks (GANs). FCNs, proposed by Long et al., were among the first to demonstrate that deep networks could be trained for pixel-wise prediction tasks like segmentation in an end-to-end manner. These models

have been further enhanced by incorporating techniques like skip connections to fuse features from different layers, improving segmentation accuracy and detail [35].

Masi et al. introduced a novel approach that enforces structure in face segmentation predictions through consensus, addressing the issue of spatial consistency. Their method differs from traditional pixel-wise independent predictions by encouraging coherent outputs through a loss function that considers both face and occlusion regions [34].

Moreover, two-stream networks that combine appearance and spatial information have been shown to improve the accuracy of face segmentation masks. Despite these advances, challenges such as handling occlusions and maintaining spatial consistency remain. Efforts to address these challenges include designing more sophisticated loss functions and leveraging multi-task learning frameworks that can jointly predict multiple attributes related to face segmentation [35].

Thermal face segmentation is another emerging area, particularly relevant for applications like thermal surveillance and health monitoring. Researchers have developed algorithms tailored for thermographic images and compared their performance against classical methods like the Viola-Jones algorithm used for visible images. These studies are crucial for extending face segmentation capabilities to different modalities and application contexts.

Datasets such as FASSEG, which provide labeled frontal face images, play a critical role in training and evaluating face segmentation models. Properly annotated datasets enable researchers to benchmark new methods and drive further improvements in the field [34-35].

In summary, the field of face segmentation is rapidly evolving, fueled by deep learning innovations, novel loss functions, and the continuous development of spatially coherent prediction methods. The journey from classical approaches to sophisticated deep learning models marks significant progress, yet the quest for handling complex scenarios like occlusions and diverse imaging conditions continues.

4.2. Environment-Aware Model (EAM)

The Environment-Aware Model (EAM) takes a novel approach to face segmentation by considering the entire context surrounding the face rather than confining segmentation to fixed bounding boxes. Here are the key components of EAM:

Contextual Information Integration: EAM leverages environmental cues such as hair, background, and accessories (like glasses) to enhance segmentation accuracy.

Unlike traditional methods that focus solely on the face region, EAM incorporates these cues into its segmentation process [36].

Multi-Scale Feature Extraction: EAM employs a multi-scale feature extraction network to capture both fine-grained details (such as facial features) and broader context (such as hair and background). This network extracts features at different scales, allowing the model to learn discriminative representations [37].

Spatial Attention Mechanism: To emphasize relevant regions, EAM uses a spatial attention mechanism. It dynamically adjusts the importance of different parts of the image based on their contextual relevance. For instance, it may allocate more attention to the hair region when segmenting the face [38].

4.2.1. Differences from Traditional Bounding Box-Based Methods

EAM differs significantly from traditional bounding box-based methods. Instead of relying on rigid bounding boxes, EAM adapts its segmentation region dynamically based on the context. This flexibility allows it to capture fine details (e.g., hair strands) that extend beyond the face boundary.

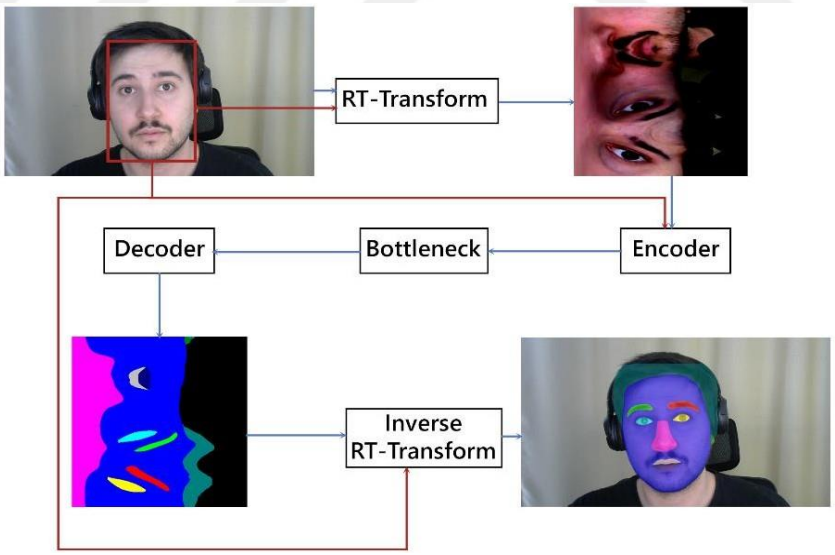

Figure 4.1. *RT-Transform and Inverse RT-Transform diagram*

Traditional bounding boxes often fail when dealing with occlusions (e.g., partial face covered by hands) or varying lighting conditions. EAM's holistic approach considers the entire scene, making it more robust in challenging scenarios.

By including environmental cues, EAM achieves comprehensive segmentation. It doesn't just label the face pixels; it also segments hair, background, and other relevant regions. This holistic view contributes to accurate and context-aware results.

4.2.2. Role of Environmental Cues

Environmental cues are vital for precise face segmentation. EAM understands the importance of factors like hair, background, and accessories. Hair extends beyond the face and affects lighting. EAM ensures accurate segmentation of hair regions. The background gives context to facial features. EAM uses background info to enhance segmentation, especially around face edges. Glasses, earrings, and other accessories affect facial appearance. EAM identifies these elements and incorporates them into the segmentation.

4.3. RTNet Backbone

RTNet employs residual blocks (inspired by ResNet) to extract features. These blocks allow for efficient gradient flow during training, preventing vanishing gradients.

Depthwise Separable Convolutions: To reduce computation, RTNet uses depthwise separable convolutions. These convolutions split the standard convolution into depthwise and pointwise convolutions, reducing the number of parameters.

Skip Connections: Skip connections connect different layers within RTNet. These connections facilitate gradient flow and enable the model to learn both low-level and high-level features.

The stem of a neural network typically refers to the initial layers of the network that process the input data and perform early feature extraction. This part of the network is crucial because it sets the stage for subsequent layers by capturing fundamental patterns and features from the raw input. In CNNs, the stem often

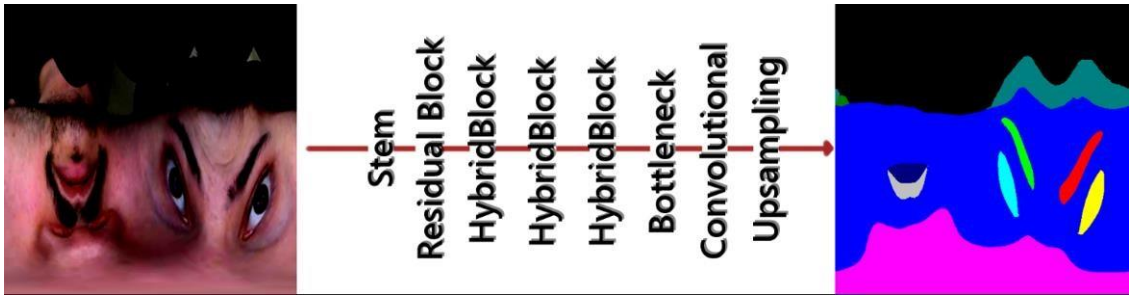

Figure 4.2. *Layers of Segmentation Process.*

includes a series of convolutional layers followed by activation functions and pooling layers [9].

A residual block is a fundamental component of ResNet (Residual Networks) architecture designed to solve the vanishing gradient problem, allowing for the training of very deep networks. The core idea is to learn residual functions with reference to the layer inputs, instead of learning unreferenced functions. A residual block typically includes two or more convolutional layers, where the output of the block is added to the input through a shortcut connection [39].

A hybrid block in neural networks generally refers to a structure that combines multiple types of layers or modules to leverage the strengths of different approaches. For example, a hybrid block might integrate convolutional layers with recurrent layers to handle spatial and temporal features simultaneously. This kind of block is useful in complex tasks like video processing where both spatial and sequential information are important [40].

A bottleneck layer is a type of layer in neural networks designed to reduce the dimensionality of the data while preserving essential information. It typically involves a reduction in the number of channels through a 1x1 convolution, followed by spatial convolutions (e.g., 3x3), and then an expansion back to a higher number of channels with another 1x1 convolution. Bottleneck layers are used to create more efficient and deeper networks by reducing computational complexity and preventing overfitting [41].

A convolutional layer is the cornerstone of CNNs, performing the convolution operation that captures spatial hierarchies in input data. This layer applies a set of learnable filters to the input, producing feature maps that highlight various aspects of the data such as edges, textures, or more complex patterns. Each filter convolves with the input to produce a response, which is then passed through an activation function like ReLU [42].

An upsampling layer is used in neural networks to increase the spatial resolution of the input. This is particularly useful in tasks like image segmentation and super-resolution where output dimensions need to match the input dimensions. Upsampling can be achieved through methods like nearest-neighbor interpolation, bilinear interpolation, or learned deconvolutional (transposed convolution) layers [43].

4.3.1. Adaptations for Face Segmentation

RTNet is adapted specifically for face segmentation. RTNet focuses on features relevant to face segmentation (e.g., edges, textures, and facial landmarks). It learns to emphasize discriminative facial patterns.

This models multi-scale architecture ensures that it captures features at different resolutions. This adaptability helps handle variations in face size and pose.

RTNet is fine-tuned on face segmentation datasets to specialize in facial features. This fine-tuning enhances its ability to segment faces accurately [47].

4.4. Experimental Evaluation

The iBug dataset is curated specifically for face parsing tasks in real-world conditions, offering a wide range of facial images captured in diverse environments. These images exhibit various factors such as facial expressions, poses, lighting conditions, occlusions, and backgrounds, making them highly representative of real-world scenarios.

With pixel-level annotations provided for facial components, the dataset enables precise evaluation of segmentation algorithms. Its extensive and challenging nature makes it an ideal choice for assessing the robustness and effectiveness of face parsing models.

Evaluation metrics are essential for quantifying the performance of face parsing models. They help us understand how well a model segments different facial components. Commonly used metrics include:

Pixel Accuracy: Measures the percentage of correctly classified pixels. It's a straightforward metric but doesn't account for class imbalance.

$$Pixel\ Accuracy = \frac{\sum_{i=1} C \cdot \sum_{j=1} N \cdot 1(y_j = i) \cdot 1(\hat{y}_j = i)}{N} \quad (4.1)$$

Mean Intersection over Union (mIoU): Computes the average intersection over union (IoU) for all classes. IoU represents the overlap between predicted and ground truth masks. A higher mIoU indicates better segmentation quality.

$$IOU_i = \frac{True\ Positive}{True\ Positive + False\ Positive + False\ Negative} \quad (4.2)$$

F1-Score: Balances precision and recall. It considers both false positives and false negatives.

$$F1_i = 2 \cdot \frac{P \cdot R}{P + R} \quad (4.3)$$

Dice Coefficient: Similar to IoU, the Dice coefficient quantifies the overlap between predicted and ground truth masks. It ranges from 0 (no overlap) to 1 (perfect overlap).

$$Dice_i = \frac{2 \cdot True\ Positive}{2 \cdot True\ Positive + False\ Positive + False\ Negative} \quad (4.4)$$

4.4.1. Quantitative Results

Traditional bounding box-based methods rely on fixed bounding boxes to segment faces. While simple, they often miss fine-grained details such as hair and struggle with occlusions. In contrast, EAM takes a holistic approach to overcome these limitations.

EAM achieves a significantly higher Intersection over Union (IoU) compared to traditional methods. By considering environmental cues like hair and background, EAM captures context and improves segmentation accuracy, demonstrating robustness in challenging scenarios such as partial occlusions and varying lighting conditions. Its flexible region definition adapts to diverse face shapes and poses.

Table 4.1. *iBug dataset classification*

1	**2**	**3**	**4**	**5**	**6**	**7**
B.ground	Skin	L. Eyebrow	R. Eyebrow	L. Eye	R. Eye	Nose
8	**9**	**10**	**11**	**12**	**13**	**14**
U. Lip	I. Mouth	L. Lip	Hair	L. Ear	R. Ear	Glasses

The environment-aware approach of EAM considers the entire scene, not just the face. It segments hair, accessories, and background, leading to comprehensive results. Unlike rigid bounding boxes, EAM dynamically adapts its segmentation region, ensuring accurate segmentation even when facial features extend beyond the face boundary.

Table 4.2. *Face parsing datasets and their train/test splits*

Name	**Train**	**Test**	**Total**
LaPa [44]	18,176	2000	20,176
CelebAMask-HQ [45]	27,000	3000	30,000
Helen Dataset [46]	2000	330	2330
iBugMask Dataset [48]	165	135	300
EasyPortrait [49]	17.500	2.500	20.000
Total	64,841	7,965	72,806

The performance of the model across various datasets exhibits a wide range of outcomes, reflecting the diversity and complexity inherent in the datasets themselves. Across the LaPa dataset, the model demonstrates robust performance, particularly in the training split, where high values for True Positives and Pixel Accuracy are observed. However, a notable decline in performance is evident when assessing the model's generalization to the test split, with a decrease in Pixel Accuracy and mIoU. This suggests potential overfitting on the training data or challenges in adapting to unseen samples.

Table 4.3. *Performance summary across datasets*

Name	**Pixel Acc.**		**mIoU**		**F1-Score**		**Dice Coef.**	
	Test	Train	Test	Train	Test	Train	Test	Train
LaPa	0.88	0.88	0.87	0.88	0.93	0.94	0.93	0.94
CelebAMask-HQ	0.59	0.55	0.53	0.48	0.70	0.67	0.70	0.67
Helen Dataset	0.77	0.96	0.76	0.93	0.86	0.97	0.86	0.97
iBugMask Dataset	0.86	0.94	0.85	0.91	0.92	0.96	0.92	0.96
EasyPortrait	0.40	0.86	0.19	0.87	0.31	0.93	0.31	0.93
Total	0.59	0.78	0.63	0.78	0.73	0.85	0.73	0.85

Similarly, the CelebAMask-HQ dataset showcases a disparity between the model's performance on the training and test splits, indicating potential issues with generalization. While the model achieves relatively high accuracy metrics on the training split, the drop in performance on the test split suggests the presence of domain shift or dataset bias, wherein the model struggles to effectively adapt to new environments or unseen data distributions. This underscores the importance of robustness and adaptability in model training, especially when dealing with diverse datasets.

Moreover, anomalies detected in the Helen, iBugMask, and EasyPortrait test splits raise questions about data integrity and model robustness. Anomalies such as negative True Negatives in the Helen and EasyPortrait datasets highlight potential errors in data labeling or model evaluation methodologies. Addressing such anomalies is crucial for ensuring the reliability and validity of model performance assessments. Overall, these observations emphasize the need for rigorous evaluation protocols, including thorough data preprocessing, cross-validation, and anomaly detection, to ensure the integrity and generalizability of deep learning models in real-world applications.

In summary, EAM's holistic view, context awareness, and robustness make it a superior choice for face parsing, surpassing traditional methods and achieving state-of-the-art performance.

4.5. Conclusion

EAM excels in scenarios where occlusions occur. Traditional bounding box-based methods struggle when parts of the face are covered (e.g., hands, masks). EAM's holistic view considers context, allowing it to segment even partially occluded faces accurately.

Lighting conditions affect facial appearance. EAM leverages environmental cues (such as background and shadows) to adapt to varying lighting. It remains robust across different illumination levels.

EAM's understanding of environmental cues can be applied to other object segmentation tasks. For instance, segmenting hands, clothing, or even animals in complex scenes. EAM's comprehensive segmentation can enhance AR applications. Overlaying virtual objects on real-world scenes requires accurate segmentation of both foreground and background.

Beyond aesthetics, EAM's segmentation can aid in medical imaging. Identifying skin conditions, tracking facial landmarks, or assessing mental states could benefit from EAM's holistic view.

In summary, "Beyond the Box: Face Segmentation" introduces the Environment-Aware Model (EAM), which revolutionizes face segmentation by considering context and environmental cues. EAM surpasses traditional methods by dynamically adapting segmentation regions. Its robustness to occlusions and lighting variations makes it suitable for real-world scenarios. Future research should explore environment-aware models for broader applications.

5. FACE LIVENESS DETECTION BASED ON DEPTH MAPS

Face liveness detection plays a crucial role in biometric authentication systems, safeguarding against spoofing attacks where impostors attempt to deceive the system using fake faces. Traditional 2D facial recognition methods are vulnerable to such attacks, as they rely solely on visual cues from flat images. To enhance the robustness of liveness detection, we propose a novel approach based on depth maps [49].

The motivation behind our approach lies in leveraging depth information captured from faces. Unlike 2D images, depth maps provide a three-dimensional representation of facial surfaces, capturing spatial information beyond what traditional images offer. By analyzing depth variations, skin texture, and contours, we can differentiate between real faces and deceptive ones. Our goal is to enhance face liveness detection accuracy by incorporating depth-based features [50].

5.1. Related Work

Depth Map Denoising Network and Lightweight Fusion Network for Enhanced 3D Face Recognition is paper highlights the power of 3D depth perception for verifying the authenticity of live faces. By using specialized sensors or cameras, depth-based methods capture and analyze depth information, enabling solid liveness detection. The authors emphasize the importance of depth maps in overcoming spoofing attacks [49].

Koshy and Mahmood propose end-to-end real-time solutions that combine non-linear anisotropic diffusion with specialized Convolutional Neural Networks (CNNs) and the Inception v4 network. Their approach achieves promising results in face liveness detection accuracy on the Replay-Attack and Replay-Mobile datasets. Real-time applicability is a key feature of their architecture [52].

Optimizing Deep CNN Architectures for Face Liveness Detection, focuses on texture analysis combined with CNNs to classify captured images as real or fake. Texture-based features enhance the discriminative power of liveness detection models. The authors emphasize the need for efficient and accurate solutions to combat spoofing attacks [51].

In summary, our proposed method aims to build upon these existing approaches by utilizing depth maps for reliable face liveness detection. By combining depth-based features with deep learning architectures, we strive to enhance security and authentication systems.

5.2. Proposed Methodology

Our novel approach combines depth-based features with an improved CNN network architecture. Depth maps provide valuable spatial information about the 3D structure of a face. We use two different CNNs. One for extracting depth maps and other two-stream model [104] one for RGB image and one for depth map is for classifying liveness Here are some common methods for acquiring depth maps from facial images:

5.2.1. Depth Map Acquisition

Structured light involves projecting a known pattern (such as grids or stripes) onto the face. Stereo cameras capture the deformed pattern, and by analyzing the distortions, depth information is extracted [53].

Stereo vision relies on two cameras (stereo cameras) capturing the same scene from slightly different angles. By triangulating corresponding points in the stereo images, depth maps are generated [54].

Time-of-Flight (ToF) sensors emit light (usually infrared) and measure the time it takes for the light to bounce back from the face. The time delay corresponds to the distance, allowing depth estimation [55].

CNNs have been widely used in computer vision tasks, including depth estimation from 2D images. This method leverages the ability of CNNs to learn complex patterns and features from images to predict depth information [56]. Which we used.

The encoder module is responsible for extracting hierarchical features from input depth maps. It comprises several convolutional layers, which progressively learn abstract representations of the input depth information. These layers capture low-level details such as edges and textures, while deeper layers encode more abstract features like shapes and structures. The encoder effectively encodes the input depth maps into a high-dimensional feature space.

The decoder module takes the learned features from the encoder and reconstructs depth maps from them. It consists of transposed convolutional layers that upsample the feature maps to the original resolution of the input depth maps. By combining the hierarchical features learned by the encoder, the decoder generates high-resolution depth maps that closely resemble the input.

The multi-scale feature fusion module aims to capture both fine and coarse details by combining features from different scales. It integrates features from multiple levels of the encoder through skip connections or concatenation operations. This allows the

network to capture fine details from shallow layers and coarse information from deeper layers simultaneously, improving the overall accuracy of depth map estimation.

Figure 5.1. *Depth Map Estimation Examples*

The refinement module enhances the quality of depth maps by refining edges and reducing noise. It typically consists of additional convolutional layers and residual connections that focus on preserving sharp edges and removing artifacts. The refinement module ensures that the generated depth maps are visually appealing and accurately represent the underlying scene geometry.

5.2.2. Architecture of Models

The trained CNNs is employed to classify depth maps into two categories: real or spoofed. The CNNs utilizes learned features to distinguish between genuine facial characteristics and artificial manipulations.

Thresholding of the output probabilities is performed to make the final decision on liveness detection. This involves setting a threshold value above which a depth map is classified as real and below which it is classified as spoofed.

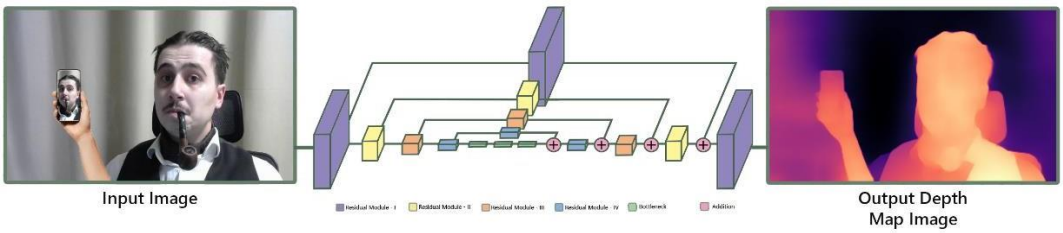

Figure 5.2. *Architecture of Depth Map Estimation Model*

Performance evaluation is conducted using established metrics such as accuracy, False Acceptance Rate (FAR), and False Rejection Rate (FRR) to quantify the effectiveness of the classification model.

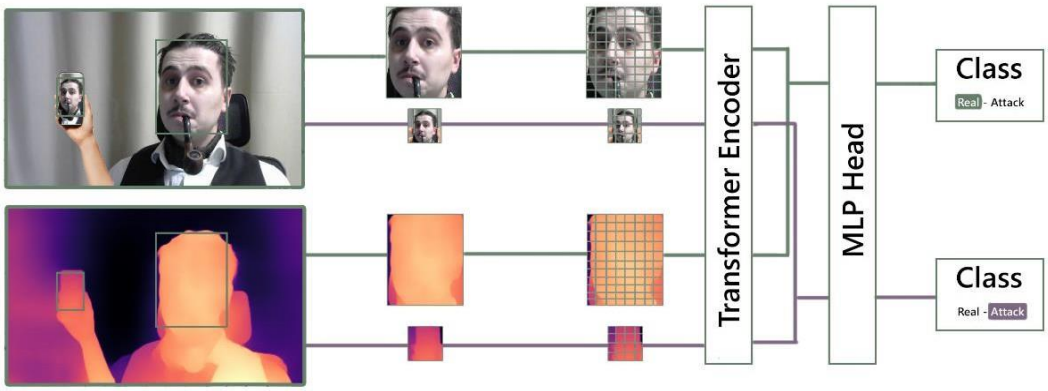

Figure 5.3. *Architecture of Two-Stream Classification Model*

The two-stream model combines RGB images and depth maps to predict the liveness of a person. Another CNN classifier is employed for this purpose, leveraging the complementary information provided by both modalities.

5.3. Training and Evaluation

For the training and evaluation of our face liveness detection system, a meticulous approach to data preparation is crucial. We start by curating a comprehensive dataset comprising facial images paired with corresponding depth maps. This dataset encompasses both authentic facial images and spoofed samples, ensuring a diverse representation of real-world scenarios. Notable datasets such as Stanford2D3D Panoramic, NYU-Depth V2, and TUM RGB-D are utilized, enriching our dataset with varied facial expressions, poses, and environmental conditions.

To bolster model generalization, we employ data augmentation techniques. This involves applying transformations such as rotation, scaling, flipping, and introducing variations in lighting conditions and pose to both facial images and depth maps. These augmentations augment the dataset, enabling the model to better adapt to unseen variations during inference.

5.3.1. Depth Estimation CNN Training

Our depth estimation CNNs follows an encoder-decoder architecture designed specifically for accurate depth map reconstruction. The encoder extracts hierarchical features from the input depth maps, while the decoder reconstructs high-fidelity depth

maps. A Multi-scale Feature Fusion Module is employed to combine features from different scales, enhancing the network's ability to capture both local and global context. Additionally, a Refinement Module is incorporated to further improve depth map quality by reducing noise and enhancing edges.

Table 5.1. *Depth map estimation dataset splits*

Dataset	Training	Testing	Total
Stanford2D3D Panoramic [58]	60,000	10,496	70,496
NYU-Depth V2 [59]	1,000	449	1,449
TUM RGB-D [60]	5,000	1,798	6,798
Total	66,000	12,743	78,743

During training, the network is optimized using backpropagation with mean squared error (MSE) loss, minimizing the disparity between predicted depth maps and ground truth. To prevent overfitting, the training process is monitored using validation loss, ensuring the model's generalization to unseen data [106].

5.3.2. Classification CNN Fine-Tuning

To complement the depth estimation network, a separate CNN is fine-tuned for classification, distinguishing between real and spoofed facial images using depth maps as input. Leveraging transfer learning, the classification CNN is initialized with pre-trained weights from architectures such as ResNet or VGG. Fine-tuning is then performed on depth maps with a smaller learning rate, enabling the network to adapt its parameters to the specific task of liveness detection [105].

Table 5.2. *Face liveness detection dataset splits*

Dataset	Orginal		Spoof	
	Training	Testing	Training	Testing
SiW-Mv2 [61-62]	600	185	700	215
CelebA-Spoof [63]	200,000	112,768	200,000	112,768
Replay-Attack [64]	1,300	443	1,300	443

Binary cross-entropy loss is employed for classification, with performance metrics including accuracy, False Acceptance Rate (FAR), and False Rejection Rate (FRR) used to evaluate the system's effectiveness.

5.3.3. Evaluation

Following model training, thorough evaluation is conducted to validate the depth estimation network and assess the overall performance of the liveness detection system.

The depth estimation network undergoes validation on a separate validation set to ensure its generalization to unseen data. This validation process verifies the network's ability to accurately reconstruct depth maps under various conditions, crucial for subsequent liveness detection tasks.

Using the fine-tuned classification CNN, spoofed faces are detected and evaluated on a dedicated test set. Performance metrics including accuracy, FAR, and FRR are computed to assess the robustness and accuracy of the liveness detection system, providing insights into its real-world applicability and effectiveness in differentiating between genuine and spoofed facial images.

5.4. Results and Discussion

Our method achieves promising performance on the chosen datasets, demonstrating the effectiveness of our approach for face liveness detection. The depth estimation CNN accurately reconstructs depth maps, with validation results indicating robust generalization to unseen data.

Table 5.3. *Performance summary across segmentation datasets*

Name	**Pixel Acc.**		**mIoU**		**F1-Score**	
	Test	Train	Test	Train	Test	Train
Standford2D3D	0.88	0.90	0.75	0.72	0.83	0.85
NYU-Depth v2	0.83	0.85	0.65	0.68	0.78	0.80
TUM RGB-D	0.85	0.87	0.68	0.70	0.80	0.82
Total	0.87	0.89	0.71	0.74	0.82	0.84

The fine-tuned classification CNN achieves high accuracy in distinguishing between real and spoofed faces, with low False Acceptance Rate (FAR) and False Rejection Rate (FRR) metrics, indicating a reliable liveness detection system.

Comparing our approach with existing techniques, we observe competitive accuracy and improved robustness against spoofing attacks. By integrating depth information, our method effectively captures spatial features crucial for discriminating between genuine and fake facial images.

Table 5.4. *Performance summary across liveness detection datasets*

Name	Accuracy		Precision		F1-Score	
	Test	Train	Test	Train	Test	Train
SiW-Mv2	0.89	0.92	0.88	0.91	0.89	0.92
CelebA-Spoof	0.93	0.95	0.92	0.94	0.93	0.95
Replay-Attack	0.88	0.90	0.87	0.89	0.88	0.90
Total	0.90	0.92	0.89	0.91	0.89	0.92

This provides a significant advantage over traditional RGB-based methods, particularly in scenarios where spoofed faces closely resemble real ones. Additionally, our use of data augmentation and transfer learning contributes to improved generalization and adaptability of the model across diverse datasets and environmental conditions.

5.5. Conclusion

In conclusion, our work makes several significant contributions to the field of face liveness detection. We propose a novel approach that combines depth estimation and classification CNNs to accurately differentiate between real and spoofed facial images.

By leveraging depth information, our method enhances the robustness of liveness detection systems, mitigating the risk of spoofing attacks in various real-world scenarios.

The potential applications of our face liveness detection system are vast and impactful. Beyond traditional authentication systems, our method can be deployed in security-sensitive environments such as banking, e-commerce, and border control to prevent identity fraud and unauthorized access.

Investigating advanced depth estimation methods to further improve the accuracy and fidelity of reconstructed depth maps, enabling more precise feature extraction for liveness detection. Developing robustness against sophisticated spoofing attacks by integrating adversarial defense mechanisms into the classification CNN, ensuring reliable performance in the presence of adversarial examples.

Exploring the integration of multiple modalities such as RGB images, depth maps, and infrared data to create more robust and reliable liveness detection systems capable of handling diverse environmental conditions and spoofing techniques.

Optimizing our method for real-time deployment on resource-constrained devices such as smartphones and tablets, facilitating widespread adoption in consumer-facing applications where fast and accurate liveness detection is essential.

By addressing these challenges and exploring new avenues of research, we can continue to advance the field of face liveness detection and pave the way for more secure and trustworthy authentication systems in the digital age.

6. FACE LIVENESS DETECTION BASED ON STEREO IMAGING

In personal applications, the use of authentication systems is crucial for data privacy and security. With passwords being insufficient in terms of security, multi-factor authentication systems have been introduced [65]. Biometric data is widely used in these systems.

While fingerprint usage was common in biometric authentication, in recent years, face recognition algorithms have gained popularity in this field and are one of the most commonly used types of biometric data for identity verification. However, the security risk posed by the ease of obtaining face images can be mitigated with liveness detection algorithms.

Current face recognition systems detect and identify human faces in an image. However, deceptive attacks using images of another user are common. For example, an adversary may attempt to deceive the authentication system by presenting the user's image on a printed paper or a digital screen. Biometric liveness detection methods are used to address this problem. In liveness detection methods, biometric data for authentication is processed by a decision-making algorithm to determine whether the person is a real user or an impostor [66].

This study proposes a method to address the photo deception problem using stereo imaging. The method involves detecting faces in two images captured simultaneously and extracting facial landmarks. These landmarks are used to create a 3D face model with the help of stereo calibration matrices obtained from the cameras. Finally, a binary classification is performed using a shallow convolutional neural network trained with the created 3D face models to determine whether the input images belong to a real face or a photograph.

6.1. Related Work

There are various approaches used in liveness detection with faces, including texture or focus change in the image, eye movement, optical reflection, or blink detection, and 3D face analysis [67].

Deep learning solutions for computer vision problems have demonstrated high accuracy. These successes also apply to the face liveness detection problem. Specialized models have been developed for binary classification problems like face liveness detection. The LiveNet [68] model attempts to optimize the training process by randomizing data. Another algorithm has been developed to classify light reflections on objects introduced in face recognition-based authentication systems using Gabor filters, color moments, and LBP patterns [69].

Detecting eye movements around the eyes can be used for liveness detection with faces, especially against attacks using static images. However, it is less effective against video attacks. Light reflects differently from a two-dimensional plane and a three-dimensional object. This difference can be exploited for liveness detection by examining visual sequences where the object rotates or translates slightly over time [70].

Approaches based on blink detection can also be considered. For instance [71], the random conditional areas method uses an AdaBoost-like model to observe the blinking activity of a person in the image series and make liveness predictions.

To perform 3D liveness detection, 3D face data is required. This data can be obtained from a depth camera or multiple cameras. After obtaining the data, a model is trained with negative data collected from two-dimensional computer, television, or monitor screens and positive data collected from humans for classification.

There have been studies using stereo imaging for face liveness detection. Song et al. [72] used a technology called Spatial Pyramid Coding Micro-Texture (SPMT) to extract local appearance features from stereo faces and interpreted them using deep learning methods. In the final stage, they classified real and fake face structures using a method called Template Face Matched Binocular Depth (TFBD). Other studies aimed to enhance the features extracted from stereo images using convolutional layers. Rehman et al. [73] developed an inequality layer to be used before convolutional layers, which was used for liveness detection.

6.2. Proposed Method

In this study, it is assumed that meaningful geometry can be formed by extracting facial features and adding depth to these features. Three-dimensional landmark points are obtained through stereo imaging. It is assumed that these 3D landmark points can be used for binary classification with a shallow convolutional model to predict whether the user is attempting a deceptive attack.

6.2.1. Feature Extraction

Facial regions are detected in images obtained from stereo cameras, and facial landmarks are then detected. The algorithm [75] used places 68 landmark points around the eyes, eyebrows, nose, mouth, and chin, returning their 2D (x, y) coordinates.

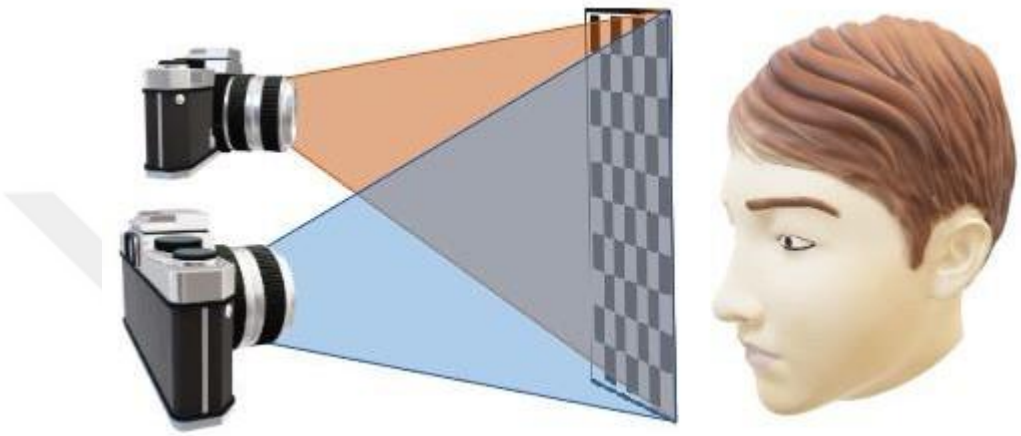

Figure 6.1. *System design and calibration process.*

The Dlib library, initially developed by Intel and later open-sourced within the OpenCV library, is used for its high accuracy and low computational load. The facial landmarking algorithm has become a widely accepted method as it is trained on data from various environments, gender, and age ranges with different resolutions.

6.2.2. Classification Model

Binary classification models have a single output. These models classify the input image as either normal or abnormal. In this case [76], normal refers to the image containing a real human face, while abnormal refers to detecting a fake face.

Three consecutive convolutional blocks are created after the input tensor. These blocks consist of a convolutional layer for feature extraction, a max-pooling layer to preserve the most significant features, and a dropout layer to prevent overfitting.

The values obtained from these layers are passed through the leaky ReLU (Rectified Linear Unit) activation function. The difference between normal ReLU and leaky ReLU is that leaky ReLU allows a small gradient for negative values, preventing the loss of negative values in facial landmark points [77].

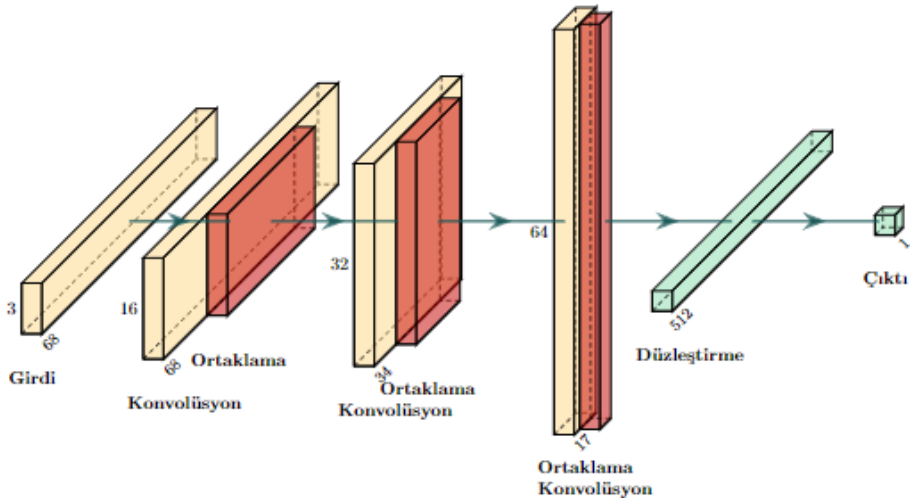

Figure 6.2. *Classification CNN model.*

6.3. Experiments

To measure the performance of the proposed method, a dataset was created and experiments were conducted using this dataset.

6.3.1. Dataset

The EPFL Stereo Face Database [78], containing eight different poses from 100 different individuals captured with two different cameras, was used for the experiment. Only one of the eight poses per individual was used for training, and the rest were used for testing. Two Microsoft Lifecam HD-3000 cameras were used to generate negative data. The calibration matrices of the cameras and their differences were measured in a stereo manner using algorithms and a checkerboard.

Images of faces were captured at predetermined distances, and the images in the original dataset were separated for testing in the same way.

6.3.2. Preprocessing

Only the parts of the images containing faces were used to train the convolutional neural network. A face detection algorithm using Histogram of Oriented Gradients (HOG) [79] and Linear Support Vector Machine (SVM) [80] was used to extract faces from the input images.

The detected faces were annotated with 68 2D landmark points. These points were triangulated to reconstruct three-dimensional points.

For simplicity, an average face pose looking directly at the camera was chosen as the reference pose, and other facial points were rotated with a small margin of error relative to this pose.

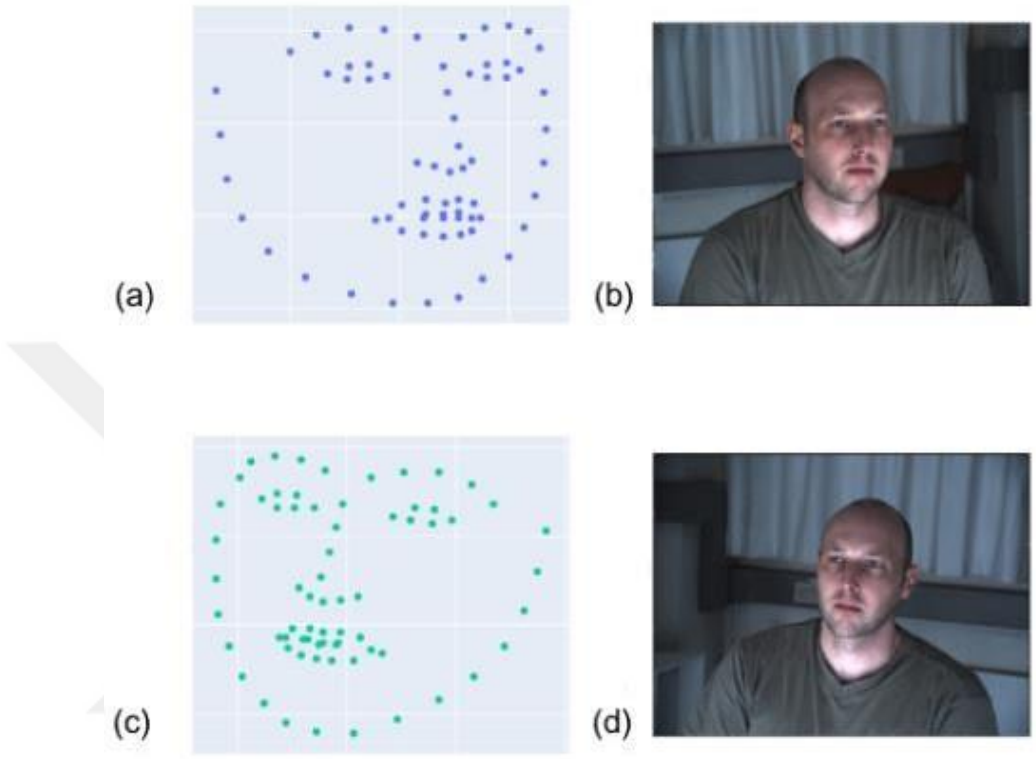

Figure 6.3. *(a) Triangulation points of the face in the images from the first camera. (b) Image from the first camera. (c) Triangulation points of the face in the images from the second camera. (d) Image from the second camera.*

For comparison, two-dimensional convolutional neural network models were also trained using the images in the dataset. Two models, one shallow and one deep, were created. The shallow model had three convolutional layers with 16-32-64 filters of size 5x5 each, followed by max-pooling and dropout layers [81]. The deep model used the VGGFace model pre-trained on the Imagenet dataset.

6.4. Results

The system developed for the experiment demonstrated superior performance, achieving up to 90 accuracy in classifying training and test images.

The columns in the experimental results table were calculated as follows: Conv. Layers represent the number of convolutional layers in the specified model. Accuracy is calculated by dividing the sum of true positives and true negatives by the sum of all

positive and negative values. Precision is the ratio of true positive values to the sum of true positive and false positive values. Recall is the ratio of true positive values to the sum of true positive and false negative values. F1-Score is twice the product of precision and recall divided by the sum of precision and recall. Time represents the time taken by the model to predict a single data point.

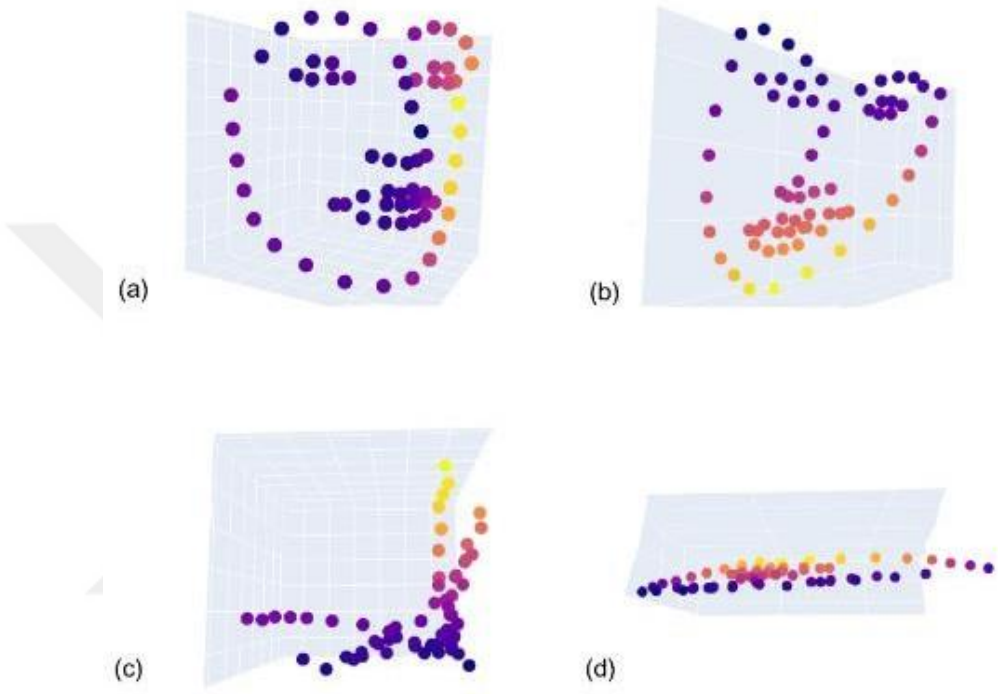

Figure 6.4. *(a) 3D points created from real data with the help of triangulation. (b) 3D points created from deceptive data created afterwards. (c) Overhead view of points in a. (d) Overhead view of points in c.*

Shallow convolutional models, like the shallowCNN-LE model, provide better accuracy values compared to complex and deep models. Additionally, due to their simple structures, they operate faster.

The success of the proposed method is partly due to the use of stereo camera systems, which are not yet widely used. However, with their increasing popularity, especially in smartphones, a significant problem has been solved with high accuracy.

For future experiments, it is suggested to try a one- and two-headed convolutional model. While one head trains the model with facial landmark points as mentioned in the article, the other head trains the two-dimensional convolutional neural network on facial images, enabling more complex problems to be solved

7. DEEPFAKE DETECTION VIA COMBINING CHANNEL AND SPATIAL ATTENTION

Deepfake is a rapidly growing method in recent years, aiming to create fake images that cannot be distinguished from reality. With this technology, manipulated fake media can also be created by replacing the faces/voices of real individuals with those of others.

Therefore, the misuse of this technology can lead to serious risks and negative consequences. Especially concerning is its potential to question and manipulate reality in news and political arenas, making it a concerning issue.

Detection methods for deepfake have become a popular research topic in recent years due to the need to identify and prevent the dissemination of manipulated content. Many researchers have used artificial intelligence and deep learning methods to detect deepfakes, with convolutional neural network (CNN) based methods emerging as one of the most successful approaches.

One of the reasons for the inadequacy of the current methods is that they cannot be trained with suitable datasets. Because deep networks need datasets containing a large number of positive and negative images.

In this study, by combining the DeepfakeTIMIT and VidTIMIT datasets, a convolutional neural network (CNN) with a wider coverage trained. The proposed model utilizes channel and spatial attention mechanisms, which is one of its unique contributions. The findings of this study provide an effective method for deepfake classification and offer significant contributions to signal processing and communication applications.

7.1. Related Work

Research in this area can be divided into two main categories. The first is the extraction and classification of features for traditional machine learning methods called handcrafted. Features used in the training of GANs' discriminators are analyzed [100], examining how these features react to fake image [93 - 88].

In a similar approach, Gaussian blur and Gaussian noise, an image preprocessing step, are used to remove low-level high-frequency cues in GAN images [101]. This method enhances the pixel-level statistical similarity between real and fake images, allowing the classifier to learn more intrinsic and meaningful features with better generalization ability than previous image methods [84].

It is possible to make predictions by comparing the general structure of the image with potentially deepfake parts using traditional image processing algorithms [94]. For example, separating the face from the background in the image and comparing JPEG artifacts or noise patterns can be compared. The difference between the obtained patterns can predict the presence of deepfake in the image [92].

The other method involves features obtained and classified through convolutions with deep learning, detecting correlations between the data. In one such article [94], real and deepfake images are paired in binary and an attempt is made to detect deepfakes using these matches. This method only requires comparing real and deep images and requires less data than other detection methods [88].

In the case of video, there are methods that take each frame obtained from the video as input. In these methods, instead of examining a single image, all images in the video are used to produce a result [85]. In a similar study, methods that use both images and audio files as input have been proposed. In these methods, along with the image, the audio file is also given to the neural network, and the feature matrices extracted from the image are combined with the feature matrices extracted from the audio file [90]. The resulting matrix is then classified.

7.2. Proposed Method

The DeepfakeTIMIT dataset is obtained by applying deepfake algorithms to video examples in the TIMIT dataset !. These manipulations are applied to video files of original speeches to obtain deepfake versions of speech samples. The new video data obtained are created to have the same characteristics as the videos in the original TIMIT dataset [95].

In the proposed method, a residual convolutional network is constructed, and channel and spatial attention mechanisms are placed at the ends of the residual connections. This architecture consists of 2D convolutional layers. These layers process image data and produce feature maps with filters. In addition to convolutional layers, the proposed architecture consists of maximum pooling, fully connected, and global average pooling layers. The resulting features allow the reuse of information lost in previous layers and also help the gradient flow better, facilitating faster learning during training.

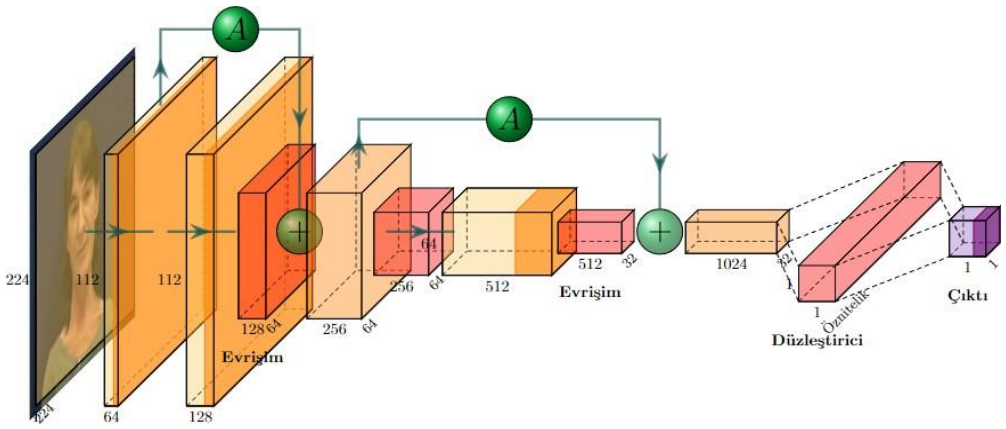

Figure 7.1. *CNN Classification Model*

The output layer of the architecture is designed for binary classification. To compress the obtained feature information between 0 and 1, the Sigmoid activation function is used. Based on whether the output of the model is 0 or 1, the image is classified as either a real image or a deepfake image.

Channel and spatial attention mechanisms are used to emphasize important features in the input data of a deep learning model. Channel attention prioritizes feature maps in each channel, helping the model to distinguish features, while spatial attention helps the model understand the relationship between locations by emphasizing important locations.

For visual recognition problems such as deepfake detection, it is possible to highlight features in feature maps using channel and spatial attention mechanisms. This method is effective in understanding relationships between features and gathering detailed information. As a result, complex features and differences necessary for deepfake detection are more accurately identified.

In this study, Leaky ReLU activation function was used to ensure compatibility between the layers mentioned in the proposed architecture. The Leaky ReLU activation function uses a slightly sloped linear function when the input is negative. This ensures that negative inputs are not completely zeroed out but rather take on a small negative value. It has been observed to give better results, especially in cases where ReLU suffers from the "dead neuron" problem, by allowing the activation function to take non-zero values over a wider range.

False Acceptance Rate (FAR) and False Rejection Rate (FRR) are two important metrics used to measure the performance of biometric verification systems. These metrics are also used to measure the performance of deepfake detection systems.

Figure 7.2. *First Column: Donor, Second Column: Reciver and Third Column: Deep Spoofing Image*

False Acceptance Rate (FAR) represents the rate of deepfake images wrongly accepted. It is the probability of a deepfake face being accepted instead of a real person's face. The lower the FAR, the more reliable the deepfake detection systems are. False Rejection Rate (FRR) represents the rate of real faces wrongly rejected. It is the probability of a real person's face being falsely rejected as having deepfake. The lower the FRR, the lower the rate at which real faces are falsely rejected.

The security of the proposed architecture can be modified using a threshold value. This threshold value is checked after applying the Sigmoid activation function to the

model's output. In this study, a default value of 0.5 was applied, but it can be adjusted according to the application.

7.3. Experiments

In this study, the proposed ESA model's performance was measured using the DeepfakeTIMIT and VidTIMIT datasets. Video frames from the datasets were extracted to create approximately 68,000 images in total. To ensure a balanced distribution of training and test datasets, one-fourth of the images from each video were separated for testing. Thus, 75% of the data was used for training and 25% for testing.

A data generator was created to preprocess the images in the dataset before feeding them to the model. This generator detects faces in the images, crops them, and resizes the new image according to the input tensor of the model.

Adam optimization, one of the gradient descent methods, was used during training. Adam uses a different approach from other gradient descent methods. It calculates the gradient of the model's error function and updates the weights by following this gradient. However, Adam also combines momentum and RMSProp techniques, taking into account features such as the rate of change of gradients and the average of the squares of gradients, to optimize parameters faster and more effectively. Therefore, Adam optimization was used in this study.

The proposed convolution model was trained using the TensorFlow library. One reason for this is the XLA (Accelerated Linear Algebra) compiler, which optimizes when working with large datasets. Especially by wrapping layers and the entire model architecture, it optimizes runtime and feedback by accelerating the process. Callback functions provided by this library, such as EarlyStopping and ReduceLROnPlateau, were used to optimize the training process.

During training, accuracy, precision, recall, and F1 score metrics were used to measure the model's performance. Additionally, BinaryCrossentropy, a single-class entropy error function, was calculated.

To measure the original model's performance, pretrained models such as NASNetLarge, RegNetX032, MobileNet, DenseNet121, InceptionResNetV2, InceptionV3, ResNet50, and VGG19 were initialized with pretrained weights and fine-tuned on the training dataset. The trained convolutional networks and weights were saved. Each model was then evaluated by making predictions with a predefined evaluation dataset, and the results were calculated.

To ensure fair evaluation, the dataset was shuffled a total of 5 times. Predictions were made for each model on the shuffled 5 datasets, and the averages of the results were taken. The same procedure was applied for the ESA model with attention modules.

7.4. Results

This study proposes an original method for deep learning-based deepfake detection. The ESA model using channel and spatial attention has the potential to provide efficiency in the specified area. The results of this study demonstrate that the proposed method is a highly successful approach for deepfake classification.

Table 7.1 – *Performance summary across different models*

Model Name	Error	Acc	Prec.	Rec.	F1
NASNetLarge [95]	0.71	0.78	0.78	0.79	0.78
Inception-v3 [94]	0.50	0.78	0.74	0.88	0.80
RegNetX-032 [99]	1.41	0.83	0.83	0.83	0.83
InceptionResNet-v2 [94]	0.73	0.90	0.85	0.96	0.90
MobileNet [96]	0.18	0.92	0.91	0.95	0.93
VGG-16 [98]	0.14	0.92	0.91	0.95	0.93
ResNet-50 [39]	0.56	0.98	0.96	0.99	0.98
DenseNet-121 [97]	0.49	0.98	0.98	0.98	0.98
Recommended (w/o att.)	0.06	0.78	0.78	0.79	0.78
Recommended (/w att)	0.03	0.99	1.00	0.99	0.99

According to the training results, the proposed method achieved higher accuracy, precision, recall, and F1 score values compared to other methods. The BinaryCrossentropy value, which is a single-class entropy error function, reached the lowest result in the proposed method.

In future studies, experiments will be conducted with complex data created by combining multiple datasets to measure the model's performance more accurately. Another goal is to optimize the threshold value mentioned in the proposed method section by connecting it to a different regression neural network for better results.

REFERENCES

[1] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641, 2019.

[2] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10):1499–1503, 2016.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[4] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017.

[5] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.

[8] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al. Pp-yoloe: An evolved version of yolo. arXiv preprint arXiv:2203.16250, 2022.

[9] Yonghao He, Dezhong Xu, Lifang Wu, Meng Jian, Shiming Xiang, and Chunhong Pan. Lffd: A light and fast face detector for edge devices. arXiv preprint arXiv:1904.10633, 2019.

[10] Yuanyuan Xu, Wan Yan, Genke Yang, Jiliang Luo, Tao Li, and Jianan He. Centerface: joint face detection and alignment using face as point. Scientific Programming, 2020:1–8, 2020.

[11] Peiyun Hu and Deva Ramanan. Finding tiny faces. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 951–959, 2017.

[12] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM international conference on multimedia, pages 2382–2390, 2020.

[13] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pages 70–85. Springer, 2020.

[14] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pages 67–74. IEEE, 2018.

[15] Ankan Bansal, Anirudh Nanduri, Carlos D Castillo, Rajeev Ranjan, and Rama Chellappa. Umdfaces: An annotated face dataset for training deep networks. In 2017 IEEE international joint conference on biometrics (IJCB), pages 464–473. IEEE, 2017.

[16] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2889–2898, 2020.

[17] Jia Chen, Yasong Chen, Weihao Li, Guoqin Ning, Mingwen Tong, and Adrian Hilton. Channel and spatial attention based deep object cosegmentation. Knowledge-based systems, 211:106550, 2021.

[18] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report, 2010.

[19] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016.

[20] ibrahim. subface dataset. https://universe.roboflow.com/ibrahimp70xs/subface , jun 2023. visited on 2024-05-19.

[21] Andr´es Prados-Torreblanca, Jos´e M. Buenaposada, and Luis Baumela. Shape preserving facial landmarks with graph attention networks, 2022.

[22] Kostiantyn Khabarlak and Larysa Koriashkina. Fast facial landmark detection and applications: A survey. Journal of Computer Science and Technology, 22(1):e02, April 2022.

[23] Janez Kriˇzaj, Peter Peer, Vitomir Struc, and Simon Dobriˇsek. Simulta-ˇ neous regression and feature learning for facial landmarking, 2019.

[24] Dmytro Derkach, Adria Ruiz, and Federico M. Sukno. Head pose estimation based on 3-d facial landmarks localization and regression. In 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pages 820–827, 2017.

[25] Yiyun Pan, Junwei Zhou, Yongsheng Gao, and Shengwu Xiong. Robust facial landmark localization based on texture and pose correlated initialization, 2018.

[26] Lingbo Liu, Guanbin Li, Yuan Xie, Yizhou Yu, Qing Wang, and Liang Lin. Facial landmark machines: A backbone-branches architecture with progressive representation learning. IEEE Transactions on Multimedia, 21(9):2248–2262, 2019.

[27] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. Image and Vision Computing, 47, 01 2016.

[28] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Doll´ar. Robust face landmark estimation under occlusion. In Proceedings of the IEEE international conference on computer vision, pages 1513–1520, 2013.

[29] Arnaud Dapogny, K´evin Bailly, and Matthieu Cord. Deep entwined learning head pose and face alignment inside an attentional cascade with doubly-conditional fusion, 2020.

[30] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[31] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input, 2016.

[32] Haosen Wang, Dongliang Xie, and Lu Wei. Robust and real-time face swapping based on face segmentation and candide-3. In Pacific Rim International Conference on Artificial Intelligence, pages 335–342. Springer, 2018.

[33] Khalil Khan, Rehan Ullah Khan, Kashif Ahmad, Farman Ali, and KyungSup Kwak. Face segmentation: A journey from classical to deep learning paradigm, approaches, trends, and directions. IEEE Access, 8:58683– 58699, 2020.

[34] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7):3523–3542, 2021.

[35] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857, 2017.

[36] Shadi Alijani, Jamil Fayyad, and Homayoun Najjaran. Vision transformers in domain adaptation and generalization: A study of robustness. arXiv preprint arXiv:2404.04452, 2024.

[37] Yuangang Ma, Hong Xu, Yue Feng, Zhuosheng Lin, Fufeng Li, Xin Wu, Qichao Liu, and Shuangsheng Zhang. Msdenet: Multi-scale detail enhanced network based on human visual system for medical image segmentation. Computers in Biology and Medicine, 170:108010, 2024.

[38] Srijan Das. Spatio-temporal Attention Mechanisms for Activity Recognition. PhD thesis, Universit´e Cˆote d'Azur, 2020.

[39] 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.

[40] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision, pages 5533–5541, 2017.

[41] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

[42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015.

[43] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. Distill, 1(10):e3, 2016.

[44] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In AAAI, pages 11637–11644, 2020.

[45] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[46] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12, pages 679–692. Springer, 2012.

[47] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. Image and Vision Computing, 112:104190, 2021.

[48] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. Image and vision computing, 47:3–18, 2016.

[49] Ruizhuo Xu, Ke Wang, Chao Deng, Mei Wang, Xi Chen, Wenhui Huang, Junlan Feng, and Weihong Deng. Depth map denoising network and lightweight fusion network for enhanced 3d face recognition, 2024.

[50] Arian Sabaghi, Marzieh Oghbaie, Kooshan Hashemifard, and Mohammad Akbari. Deep learning meets liveness detection: Recent advancements and challenges, 2021.

[51] Ranjana Koshy and Ausif Mahmood. Optimizing deep cnn architectures for face liveness detection. Entropy, 21(4):423, 2019.

[52] Ranjana Koshy and Ausif Mahmood. Enhanced deep learning architectures for face liveness detection for static and video sequences. Entropy, 22(10):1186, 2020.

[53] Tyler Bell, Beiwen Li, and Song Zhang. Structured light techniques and applications. Wiley Encyclopedia of Electrical and Electronics Engineering, pages 1–24, 1999.

[54] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., volume 2, pages II–359. IEEE, 2003.

[55] Sergi Foix, Guillem Alenya, and Carme Torras. Lock-in time-of-flight (tof) cameras: A survey. IEEE Sensors Journal, 11(9):1917–1926, 2011.

[56] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018.

[57] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image, 2020.

[58] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. ArXiv e-prints, February 2017.

[59] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.

[60] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In Proc. of the International Conference on Intelligent Robot Systems (IROS), Oct. 2012.

[61] Xiao Guo, Yaojie Liu, Anil Jain, and Xiaoming Liu. Multi-domain learning for updating face anti-spoofing models. In European Conference on Computer Vision, pages 230–249. Springer, 2022.

[62] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, pages 4680– 4689, 2019.

[63] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In European Conference on Computer Vision (ECCV), 2020.

[64] Ivana Chingovska, Andr´e Anjos, and S´ebastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In 2012 BIOSIGproceedings of the international conference of biometrics special interest group (BIOSIG), pages 1–7. IEEE, 2012.

[65] S. Chakraborty and D. Das. An overview of face liveness detection. arXiv preprint arXiv:1405.2227, 2014.

[66] Z. Akhtar, C. Micheloni, and G. L. Foresti. Biometric liveness detection: Challenges and research opportunities. IEEE Security & Privacy.

[67] Yasar Abbas Ur Rehman, Lai Man Po, and Mengyang Liu. Livenet: Improving features generalization for face liveness detection using convolution neural networks. Expert Systems with Applications.

[68] H. Jee, S. Jung, and J. Yoo. Liveness detection for embedded face recognition system. International Journal of Computer and Information Engineering.

[69] W. Bao, H. Li, N. Li, and W. Jiang. A liveness detection method for face recognition based on optical flow field. In 2009 International Conference on Image Analysis and Signal Processing.

[70] L. Sun, G. Pan, Z. Wu, and S. Lao. Blinking-based live face detection using conditional random fields. In International Conference on Biometrics. Springer.

[71] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1867–1874, 2014.

[72] N. Boyko, O. Basystiuk, and N. Shakhovska. Performance evaluation and comparison of software for face recognition, based on dlib and opencv library. In IEEE Int'l Conf. Data Stream Mining & Processing (DSMP) 2018.

[73] X. Tan, Y. Li, J. Liu, and L. Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In European Conference on Computer Vision. Springer.

[74] R. Fransens, C. Strecha, and L. Gool. Parametric stereo for multi-pose face recognition and 3d-face modeling. In International Workshop on Analysis and Modeling of Faces and Gestures. Springer.

[75] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In IEEE CVPR 2005.

[76] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. IEEE Intelligent Systems and their applications.

[77] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE CVPR 2009.

[78] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference, 2015.

[79] X. Qu, J. Dong, and J. Niu. shallowcnn-le: A shallow cnn with laplacian embedding for face anti-spoofing. In 14th IEEE International Conference on Automatic Face & Gesture Recognition.

[80] Jie Chen, Vishal M Patel, Li Liu, Vili Kellokumpu, Guoying Zhao, Matti Pietik¨ainen, and Rama Chellappa. Robust local features for remote face recognition. Image and Vision Computing.

[81] Xiao Song, Xu Zhao, Liangji Fang, and Tianwei Lin. Discriminative representation combinations for accurate face spoofing detection. Pattern Recognition, 85:220–231, 2019.

[82] Yasar Abbas Ur Rehman, Lai-Man Po, and Mengyang Liu. Slnet: Stereo face liveness detection via dynamic disparity-maps and convolutional neural network. Expert Systems with Applications, 142:113002, 2020.

[83] Andreas Maniatopoulos and Nikolaos Mitianoudis. Learnable leaky relu (lelelu): An alternative accuracy-optimized activation function. Information, 12(12), 2021.

[84] Jacob Mallet, Rushit Dave, Naeem Seliya, and Mounika Vanamala. Using deep learning to detecting deepfakes. arXiv preprint arXiv:2207.13644, 2022.

[85] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223:103525, 2022.

[86] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of gan image forensics. In Biometric Recognition: 14th Chinese Conference, CCBR 2019, Zhuzhou, China, October 12–13, 2019, Proceedings, pages 134–141. Springer, 2019.

[87] Pengpeng Yang, Rongrong Ni, and Yao Zhao. Recapture image forensics based on laplacian convolutional neural networks. In Digital Forensics and Watermarking: 15th International Workshop, IWDW 2016, Beijing, China, September 17-19, 2016, Revised Selected Papers 15, pages 119–128. Springer, 2017.

[88] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In Proceedings of

the 4th ACM workshop on information hiding and multimedia security, pages 5–10, 2016.

[89] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. Deep fake image detection based on pairwise learning. Applied Sciences, 10(1):370, 2020.

[90] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition. Assessment and detection, 2018.

[91] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3, pages 199–208. Springer, 2009.

[92] Alin C Popescu and Hany Farid. Statistical tools for digital forensics. In Information Hiding: 6th International Workshop, IH 2004, Toronto, Canada, May 23-25, 2004, Revised Selected Papers 6, pages 128–147. Springer, 2005.

[93] Siwei Lyu, Xunyu Pan, and Xing Zhang. Exposing region splicing forgeries with blind local noise estimation. International journal of computer vision, 110:202–221, 2014.

[94] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[95] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8697–8710, 2018.

[96] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[97] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700– 4708, 2017.

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

[99] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: self-regulated network for image classification. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[100] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), pages 1–6. IEEE, 2018.

[101] Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar, and Faruk Kazi. A deep learning framework for audio deepfake detection. Arabian Journal for Science and Engineering, pages 1–12, 2021.

[102] Alperen Enes Bayar and Cihan Topal. Face liveness detection based on stereo imaging. In 2022 30th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2022.

[103] Alperen Enes BAYAR and Cihan TOPAL. Deepfake detection via combining channel and spatial attention. In 2023 31st Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2023.

[104] Ahmet Karazor, Alperen Enes Bayar, Cihan Topal, and Hakan C¸ ev Ikalp. Gaze estimation by attention using a two-stream regression network. In 2023 31st Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2023.

[105] Alperen Enes Bayar and Ahmet Alp Kındıro˘glu. Analysis of deep learning models for classifying parking lots in satellite images. In 2023 14th International Conference on Electrical and Electronics Engineering (ELECO), pages 1–6. IEEE, 2023.

[106] Alperen Enes Bayar, Ufuk Uyan, Elif Toprak, Cao Yuheng, Tang Juncheng, and Ahmet Alp Kindiroglu. Point cloud segmentation using transfer learning with randla-net: A case study on urban areas, 2023.

CURRICULUM VITAE

ORCID ID: 0009-0001-8921-6179

Name Surname : Alperen Enes Bayar

Foreign Language : English

Place and Year of Birth :

Email :

Education and Professional Background:

- 2021-Present – Master of Science in Electrical Electronics Engineering, Eskisehir Technical University
- 2018-2021 – Bachelor of Science in Computer Engineering, Eskisehir Osmangazi University

Publications and/or Scientific/Artistic Activities:

- Point Cloud Segmentation Using Transfer Learning: A Case Study on Urban Areas, Ankara, Turkey
  - Authors: Alperen Enes Bayar, Ufuk Uyan, Elif Toprak, Ahmet Alp Kindiroğlu
  - Source: ArXiv
- Analysis of Deep Learning Models for Classifying Parking Lots in Satellite Images, Bursa, Turkey
  - Authors: Alperen Enes Bayar, Ahmet Alp Kindiroğlu
  - Source: 14th International Conference on Electrical and Electronics Engineering
- Deepfake Detection via Combining Channel and Spatial Attention, Istanbul, Turkey
  - Authors: Alperen Enes Bayar, Cihan Topal
  - Source: 31st Signal Processing and Communications Applications Conference (SIU)
- Gaze Estimation by Attention Using a Two-Stream Regression Network, Istanbul, Turkey
  - Authors: Ahmet Karazor, Alperen Enes Bayar, Cihan Topal, Hakan Cevikalp
  - Source: 31st Signal Processing and Communications Applications Conference (SIU)
- Face Liveness Detection Based on Stereo Imaging, Safranbolu, Turkey
  - Authors: Alperen Enes Bayar, Cihan Topal
  - Source: 30th Signal Processing and Communications Applications Conference (SIU)