

From Pavements to Patterns: Unravelling Trends in Age, Gender, and Attire for Societal Awareness

SAAD BIN YOUSUF

M.Phil



THE UNIVERSITY OF
SYDNEY

Supervisor: Dr. Simon Poon

Associate Supervisor: Dr. Matloob Khushi

The research reported in this thesis was supported by the
award of a Research Training Program scholarship
to the M.phil Candidate.

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

4 October 2024

Statement of Originality

I, Saad B. Yousuf, an M.Phil student at The University of Sydney, Australia, hereby certify that the work presented in this thesis entitled “From Pavements to Patterns: Unravelling Trends in Age, Gender, and Attire for Societal Awareness” is a result of my own original research conducted under the supervision of Prof. Simon Poon and Dr. Matloob Khushi who have provided me with the necessary guidance and support throughout the duration of my study. I certify that the work presented in this thesis is my own original work and has not been taken from any other sources except where fully acknowledged.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Sign: _____

Saad Bin Yousuf

Abstract

The recent developments in computer vision and deep learning have opened a range of different application domains, such as object detection, facial recognition, medical imaging, and smart cities. The focus of this thesis is on object detection. Specifically, we investigate apparent age estimation and gender classification with apparel detection and classification.

The existing literature does not address the concurrent estimation of apparent age, gender classification, and attire detection in the wild. To address this limitation, we create a novel dataset comprising age, gender, and attire labels. After labelling the dataset, we apply face alignment and image enhancement techniques to improve the existing age estimation and gender classification algorithms' efficacy. Our experimental results show that RetinaNet has the highest precision (1.0) and recall (0.97) among the analysed algorithms for face detection on our test set. Furthermore, we observe that varying the area around the face during the training phase impacts the efficacy of the analysed algorithms. We illustrate that incorporating $\sim 40\%$ of the area around the face coupled with image enhancement via GFP-GAN maximises the efficacy of the analysed algorithms for apparent age estimation and gender classification. We show that FRCNN is the best performing algorithm for estimating ages (mean absolute error of 6.85) on our test set, whereas, DETR-Deform is the best performing algorithm for gender classification (average precision of 0.89 for female and 0.91 for male) and attire detection (mean average precision of 0.83). The experimental results demonstrate the practicality of our approach in real-world settings.

Acknowledgements

This journey would not have been possible without the guidance, support, and patience of many, and I take this opportunity to express my deepest gratitude to each of them.

I am profoundly thankful to Associate Professor Simon Poon and Dr. Matloob Khushi for their unwavering support and invaluable expertise as my supervisors. Their guidance has been instrumental in navigating the complex challenges of my research, providing me with the tools and confidence needed to pursue my studies.

Special thanks to Dr. Kalina Yacef, my research methods teacher, whose profound insights into data-driven research methodologies have significantly shaped how I approached my thesis. Her mentorship has been invaluable in helping me understand the intricacies of scientific inquiry.

I also wish to acknowledge Dr. Saad Sajid Hashmi's encouragement and insightful discussions, which have greatly enriched my research experience. I also appreciate Associate Professor Bing Zhou and Dr. Ali Anaissi's critical insights and feedback during my Annual Progress Reviews. Their contributions have been essential in refining my research direction and methodologies.

A special note of gratitude goes to Nathaniel Butterworth, in charge of High-Performance Computing, who was instrumental in implementing numerous algorithms essential to my research. His expertise and support in the utilisation of HPC resources have been invaluable.

To all my lab mates and colleagues, thank you for the stimulating discussions and the fun we have had in the last few years. Your companionship during our academic journey has been a cherished part of my experience. Lastly, my heartfelt thanks go to my family, whose love and sacrifices have been my anchor and support throughout this endeavour.

Dedication

This thesis is dedicated to the most significant people in my life: my wife and children. Your love, patience, and understanding have guided me throughout this demanding yet fulfilling journey. Your steadfast support and encouragement have helped me to remain focused and motivated. You have been my source of strength and inspiration, and I am eternally grateful to have you by my side. To my mother and siblings, thank you for instilling in me the principles of hard work and determination. Your advice and support have been invaluable to me, and I will always look up to you as my role models. I would not have achieved this milestone without your encouragement and faith in me. This thesis is also dedicated to my extended family, who have provided me with a nurturing and supportive environment. I especially feel the need to mention Muhammad Bhai, who supported me throughout my academic endeavours. I am deeply thankful for your generosity and readiness to assist, which has profoundly impacted my life. Finally, I want to extend my gratitude to everyone who has supported me in various ways throughout this journey. Your contributions and support have been essential to successfully completing this thesis. With heartfelt thanks and love, I dedicate this thesis to my wife, children, parents, siblings, and extended family.

CHAPTER 1

Authorship Attribution Statement

The following publication was produced during my M-Phil candidature. I designed the study, curated the dataset, performed experiments, and drafted the paper. Section 3.1.3 and 3.1.4 contain methods from the publication, and Section 4.2.2 contains the results from the publication.

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Sign: _____

Saad Bin Yousuf

Sign: _____

Dr. Simon Poon

- (1) **Yousuf, S. B.**, Sajid, H., Poon, S., & Khushi, M. (2019). IMDB-Attire: A Novel Dataset for Attire Detection and Localization. In Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part II 26 (pp. 543-553). Springer International Publishing.

Furthermore, the following manuscript is currently under review:

- (1) **Yousuf, S. B.**, Hashmi, S. S., Khushi, M., & Poon, S. (2025). From Pavements to Patterns: Unravelling Trends in Age, Gender, and Attire for Societal Awareness. Under review in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2025.

Contents

Statement of Originality	ii
Abstract	iii
Acknowledgements	iv
Dedication	v
Chapter 1 Authorship Attribution Statement	vi
Contents	viii
List of Figures	xi
Chapter 1 Introduction	1
1.1 Thesis Objective	2
1.2 Research Questions	4
1.2.1 RQ1: Multi-class Labelling - Multi-task Learning	4
1.2.2 RQ2: Face to Facial-Peripheral Ratio	4
1.2.3 RQ3: Image Enhancement and its Effects on Age and Gender Prediction .	5
1.3 Contributions	5
1.3.1 RQ1: Multi-class Labelling - Multi-task Learning	5
1.3.2 RQ2: Face to Facial-Peripheral Ratio	6
1.3.3 RQ3: Image Enhancement and its Effects on Age and Gender Prediction .	6
1.4 Organisation of Thesis	7
Chapter 2 Literature review	8
2.1 Background	8
2.2 Object Detection	10
2.2.1 Two Stage Detectors	13

2.2.2	One Stage Detectors	16
2.2.3	Basic Building Blocks and Techniques for Object Detection	20
2.2.4	Speeding up of Detection	26
2.3	Face Detection and Alignment Techniques	27
2.3.1	Multi-Stage Face Detectors	27
2.3.2	Two Stage - Face Detectors	28
2.3.3	One stage - Face detectors:	28
2.4	Overview of Datasets in Object Detection	30
2.4.1	Face Detection Datasets and related work	32
2.4.2	Age and Gender Datasets and influential works	34
2.4.3	Attire detection Datasets and influential works	36
Chapter 3	Methods	39
3.1	Construction of Dataset	39
3.1.1	Threats to Validity	43
3.1.2	Annotation of labels of Attire	44
3.1.3	Labelling Attire	45
3.1.4	Annotations in Detail	48
3.1.5	Annotation of Age and Gender	51
Chapter 4	Results	56
4.1	Metrics for Evaluating Object Detection	57
4.2	Dataset construction	61
4.2.1	Image Enhancement Techniques	63
4.2.2	Dataset Statistics	64
4.2.3	Face Detection Results	66
4.2.4	Face Alignment Results	68
4.2.5	Training Charts	70
4.2.6	Results of Face to Facial-Peripheral Ratio	72
4.2.7	Image Enhancement Results on Age and Gender Prediction	74
4.2.8	Age and Gender Results on state-of-the-art algorithms	76

4.2.8.1	Mean Absolute Error of Ages on facial Images	79
4.2.8.2	Gender Precision on Dataset	80
4.2.8.3	Attire Detection Results on the state-of-the-art algorithms	81
4.2.9	Applicability in Real-World Settings	85
4.2.10	Heatmaps Providing Insights	86
Chapter 5	Conclusion	89
5.1	Concluding Remarks	89
5.2	Future Work	91
Bibliography		93

List of Figures

2.1	Computer visual interpretation	10
2.2	Timeline for object detection algorithms	10
2.3	Alex-Net Architecture	11
2.4	HoG Detector Effects	13
2.5	RCNN- Architecture	14
2.6	Spatial Pyramid Network	15
2.7	Region Proposal Network	16
2.8	Approaches to Feature Pyramid	17
2.9	YOLO Vs SSD	18
2.10	Transformer Architecture	20
2.11	Deformable Parts Model	21
3.1	Process of Constructing a Dataset	40
3.2	Flow Chart of the Process Overview to Label a Dataset.	42
3.3	Overview of LableImg Interface	44
3.4	Sample of Attire Annotation	45
3.5	Sizes of Images	46
3.6	Tine Objects labelled in Dataset	48
3.7	Versatile labels of the same kind	49
3.8	Group photo highlighting diverse attire, ages, and genders detection.	50
3.9	Casual indoor scene showcasing context-aware detection of human figures.	50
3.10	Clear facial detection and labelling of age and gender in a crowded formal setting.	51
3.11	Face detection across angles and lighting, highlighting age diversity naturally.	51
3.12	Outdoor group scene showcasing diverse age and gender detection in varied lighting.	52

3.13	Outdoor scene capturing playful interaction, showcasing diverse facial expressions detection.	52
3.14	Detection of diverse genders, ages, expressions, and attire in professional settings.	53
3.15	Studio image showcasing facial detection and labeling in professional, media-related contexts.	54
4.1	Ground Truth — True Positives - False Positive	58
4.2	Intersection Over Union	59
4.3	Precision Recall Curves Comparison	60
4.4	Comparison of Image Enhancements	62
4.5	Bar graph showing statistics of Attire	65
4.6	Figure shows a comparison of face detection Algorithms	67
4.7	Face Ratio and Age Estimation Experiments	68
4.8	Training Graphs for all algorithms	70
4.9	Face to Facial-Peripheral Ratio and Age Estimation Results	72
4.10	Ranking of Algorithms based on performance	78
4.11	Average precision of different algorithms to predict the gender on our dataset.	80
4.12	Chart comparing Algorithms on Attire Detection	84
4.13	Snapshots of Implementation System in Mall Setting.	85
4.14	Heatmaps	86

CHAPTER 1

Introduction

In the domain of computer vision, a deep understanding of face detection, age estimation, gender classification, and attire detection is very important to realise real-world applications, such as security systems (Mohammed et al. 2022), personalised advertisements (ELKarazle et al. 2022; Kuprashevich and Tolstykh 2023), social media applications (Pandey 2023; Kuprashevich and Tolstykh 2023), and various visual analytics such as generating real-time data from people in the real world (ELKarazle et al. 2022).

These advanced face detection and age estimation algorithms can improve public safety measures in public spaces such as airports, train stations and bus stations. They can also monitor public events such as conferences and concerts. It can also identify missing people and monitor persons of interest (De Felice et al. 2023). If properly used, these algorithms can identify people's attributes such as age and gender, which can be used to tailor advertisements according to the customers' preferences, enhancing customer satisfaction and engagement of customers (Alhalabi et al. 2021).

Integrating attire with age and gender can help identify and provide useful insights into cultural trends and social behaviours. This can also help sociologists and market researchers study consumer behaviour and societal norms (Twigg 2007). Incorporating attire detection has the potential to revolutionise the retail and fashion industry by gathering analytics on clothing trends. This can help stores organise in a better and more effective way. This can aid inventory management, marketing and personalised shopping experiences (Abbas et al. 2024). An AI-based system that can detect faces, estimate age and gender, and identify attire in real-time can enhance surveillance systems, enabling quicker responses to potential threats and improving overall security infrastructure (Alhalabi et al. 2021; Pramilarani et al. 2024).

The field of computer vision and machine learning is progressing rapidly, and recently, the performance of deep learning algorithms has dramatically improved computer vision tasks (Huang et al. 2022). Computer science has transformed from using simple algorithms for detecting faces to leveraging deep learning algorithms and developing and refining accompanying datasets to improve the accuracy and performance of these machine learning solutions (Ghani et al. 2022). Previous work has addressed all three tasks, namely (i) age estimation (ELKarazle et al. 2022), (ii) gender classification (Kuprashevich and Tolstykh 2023), and (iii) attire detection (Pandey 2023), individually. However, to our knowledge, no dataset or study was conducted to tackle all three tasks. Furthermore, the related studies are benchmarked against their datasets produced in controlled environments (Siddiqi et al. 2022).

Age estimation, gender classification and attire detection are very complex problems. The reason is that their visual representation is very different. For example, for age and gender, we have to analyse facial features. In contrast, for attire, depending on the attire and setting, it becomes a very complex problem. Most researchers have worked on creating datasets in controlled environments or settings, which fail in the real world (Huang et al. 2022). All three problems are very complex, as the visual representation of all three is very different. Multiple studies (Geirhos et al. 2020; Bansal et al. 2021) show that the dataset is the most important piece in a deep learning solution.

This thesis aims to produce a novel approach to addressing all three tasks integrated into one deep-learning solution that can be deployed and utilised in real-world scenarios. Besides, the thesis aims to develop a deeper understanding of these technologies and enhance image analysis to provide significant theoretical insights into these algorithms and a framework for practically implementing and deploying these solutions in the real world.

1.1 Thesis Objective

Estimating age, gender, and attire simultaneously in the wild is challenging. The real world is forever changing and dynamic; computers face different scenarios, such as wilderness, urban streets, cluttered spaces and poorly lit areas. In these situations, it is very hard to

estimate age and gender, especially in real-time. If one could develop a system that could detect age, gender, and attire simultaneously in real time, the engineering application could be far-reaching. Age and gender are instrumental in identifying human beings, and attire is a cultural marker, reflecting regional norms and social statuses. Brand affiliations within clothing further delineate societal hierarchies and reflect individual attitudes toward life. This recognition holds relevance in numerous sectors, including surveillance, fashion, and retail analytics. While recent advancements in deep learning algorithms offer promise in this domain, their effectiveness depends on extensive manual labour for image annotation. In this thesis, the main problem addressed is to take a raw image, identify if humans are in it, and then estimate the age, gender, and attire of the people in the image. However, raw images can be of any type, and the people present in the images may have various orientations. The people present in the images may be one or many. The distance between the people and the camera may also vary.

The main objective of this thesis is to develop a solution to detect age, gender, and attire simultaneously. Another objective of this research is to evaluate how these object detection algorithms work on multitask computer vision problems. We also aim to analyse the effects of image quality on age and gender estimation in the wild. We also aim to gather insights on how face detection and age-gender estimation occurs in the wild. We hope this study will lay the foundations of practical engineering systems in the real world. We also aim to publish our results and findings to pave the way for further exploration into the domain.

1.2 Research Questions

This thesis is motivated by the following research questions:

- **RQ1 (Multi-class Labelling - Multi-task Learning):** Can deep learning algorithms simultaneously learn Multi-class labels of (i) age, (ii) gender, and (iii) attire and also differentiate between problem types successfully in a real-world environment?
- **RQ2 (Face to Facial-Peripheral Ratio):** How does incorporating different percentages of the area around a face impact the outcome of deep learning algorithms in predicting age and gender in the wild?
- **RQ3 (image enhancement techniques):** How do image enhancement techniques affect apparent facial age and gender estimation?

1.2.1 RQ1: Multi-class Labelling - Multi-task Learning

One of the challenges we face while training deep learning algorithms is the difference in the visual construction of different labels and the variety of multiple labels. One of the important factors in training these algorithms is the quality of data being used. Based on how complex and comprehensive the dataset is, the performance of these algorithms can vary. Designing custom datasets, developing customised loss functions, and distinguishing between multiple labels is also a complex computer vision task. Developing a carefully curated dataset is crucial in training deep learning algorithms to address real-world problems.

1.2.2 RQ2: Face to Facial-Peripheral Ratio

The second research question was to see whether incorporating different percentages of the area around faces impacts the outcomes of deep learning algorithms in predicting age and gender in the wild. While exploring the field, we came across a study (Rothe et al. 2015) which mentioned the effect of the area around the face in predicting age; however, after our initial exploration, we felt that at first, there are not many in-the-wild data sets to predict age and gender. Identifying feature relevance in which areas around the face are more informative

for future extraction to predict age and gender requires extensive experimentation. Also, ensuring that models can work with occluded faces and still predict age and gender is an unexplored research area.

1.2.3 RQ3: Image Enhancement and its Effects on Age and Gender Prediction

To our knowledge, image enhancement methods and their influence on age estimation in the wild is an unexplored domain of computer vision. The potential effects of image enhancements on age and gender estimation have not been adequately investigated. Ensuring image quality and consistency has consistently yielded improvements in the performance of deep learning models across various real-time situations; recently, this field has garnered considerable research attention. However, incorporating image enhancement techniques into deploying computationally intensive deep learning algorithms presents an additional challenge that necessitates further research and exploration.

1.3 Contributions

Addressing the three research questions posed in section 1.2, we will briefly discuss our contributions to the field of computer science. We not only provide a novel dataset to tackle the problem of simultaneous age, gender, and attire but also provide insights into the performance of some state-of-the-art object detection algorithms.

1.3.1 RQ1: Multi-class Labelling - Multi-task Learning

To address simultaneous age estimation, gender classification, and attire detection, we construct a comprehensive dataset incorporating various manually annotated attire annotations and a comprehensive annotation of faces, identifying the age and gender of different personnel in the real world (**RQ1**). Specifically, we select 8000 images from a set of 150,000 images and annotate 60 different attire labels in the images. The images in our dataset contain different

settings, such as indoors/outdoors, varied body postures, and people of varying ages. All wearable items present in the images (even finger rings) were appropriately annotated.

We also provide a comprehensive analysis using quantitative research methods and experimental modelling to address the research questions. Furthermore, we provide multiple insights into various face detection algorithms that have been applied and tested on our dataset to gauge their efficiency and precision in the wild, as addressed in Section (4.2.3). Another contribution is the benchmark of benchmark the-art algorithms on our datasets. We also compare age, gender, and attire on state-of-the-art object detection algorithms that address (RQ1).

1.3.2 RQ2: Face to Facial-Peripheral Ratio

To address the second question, we provide a comprehensive analysis by running extensive experiments on facial features. Specifically, we conduct training of algorithms for various input data by varying the area around the face. The algorithms estimate the face’s age and we identify the most suitable face-surround-area to face ratio for the best age and gender results. We also provide results from our experiments and a comprehensive sub-dataset carefully curated to address **RQ2** (*see* Section 4.2.6 for details).

1.3.3 RQ3: Image Enhancement and its Effects on Age and Gender Prediction

To address image enhancements and their effects on age and gender estimation, we used the existing work of (Lim et al. 2017) as a starting point. However, after the first few experiments, we observed that GFP-GAN (Wang et al. 2021) worked well and was better suited for our problem. We provide a comprehensive sub-dataset with image enhancement applied to the faces in the images, and we also discuss in detail the effects of image enhancements on age estimation and gender classification

Our work offers scale-able solutions that can be deployed in various applications, from security surveillance to social media platforms, enhancing user experience and safety. The findings provide a foundation for further research into integrating labels in real-world datasets, such as emotional recognition and pose estimation.

Our research is a significant step in integrating age estimation and gender classification with attire detection within a single deep-learning Framework. By providing valuable insights into these deep learning frameworks and addressing theoretical and practical challenges, this thesis contributes to advancing computer vision technologies by paving the way for further research into this field.

1.4 Organisation of Thesis

The remained of the thesis is divided as follows. Chapter 2 gives the background information and provides the literature review. We present the methodology of our approach in Chapter 3. Chapter 4 illustrates the results of our methodology. Finally, Chapter 5 concludes the thesis and discusses the future research directions.

CHAPTER 2

Literature review

2.1 Background

The thesis, “From Pavements to Patterns: Unravelling Trends in Age, Gender, and Attire for Societal Awareness”, explores the domain of computer vision and deep learning. Reflecting on human vision, we view the whole field in front of us; however, unconsciously, we filter out all the unnecessary things and only focus on the objects of interest. For example, if you are working on a computer, you will only be focused on the computer screen. Although our eyes take in visual data from all angles, our brains consciously filter out any unnecessary information. When you work on a computer, you tend to ignore the visual information around your screen. However, your brain automatically distinguishes between important and unimportant data by filtering out the unneeded information. If something is important in your peripheral vision, your brain will process it and alert you.

The quest to emulate the human brain’s intricate processing of visual stimuli has become a focal point for computer vision-based scientific endeavours. One pivotal aim is to replicate human-like visual processing artificially, aiming towards achieving the sophistication of our innate cognitive abilities. Specifically delving into age, gender, and attire classification. These attributes form the cornerstone of human identification when encountering unfamiliar individuals. However, deciphering these features in real-world settings presents a formidable challenge.

The initial phase involves identifying these attributes in an image and separating between important and peripheral elements. Subsequently, this process selectively retains the essential

facets of the image while discarding extraneous visual information. Let us look at the features related to age, gender, and attire. While the distinctions of age and gender are independent, attire and the features that help classify attire are highly correlated. Attire, when it is identified, has a strong correlation with age and gender. For example, if a random picture is identified as a 50-year-old man with a casual shirt and pants, if he changes his pants to shorts, then the same algorithm may identify him as a 30-year-old man, as shorts are more commonly associated with the youth. Similarly, if hair colour is given priority, then grey or whitish hair would classify the person as elderly as compared to brown or darker hair colours. In the same way, any facial hair detection may indicate an adult or older male.

This chapter will overview the relevant field and summarise previous research. We will begin by discussing object detection, face detection and alignment, age estimation and gender classification. Lastly, we will cover attire detection and provide details on existing datasets addressing the above-mentioned topics.

To get a firm understanding of this thesis, it is essential to have a basic understanding of what images are and how computers perceive these images. Images and video frames represent 3D scenes in the form of 2D projections. When a computer interprets an image, it perceives an array of pixel values. So, when we say that the image size is 480×360 , there are 480 rows of pixels and 360 columns of pixels in the form of a grid. To better understand this concept, let us take a random black-and-white image.

The number of distinct levels between full white and full black defines the image's bit depth. For example, in an 8-bit image, there are 256 levels between white and black. Each numerical value within this array, ranging from 0 to 255, represents the pixel's intensity at that specific location. Although these numerical values might lack inherent meaning during image classification, they constitute the sole input data accessible to the computer.

In modern video standards, the default aspect ratio (the ratio of the image to height) is 16:9, and simply indicating the height of the video frame communicates complete information about the frame unless the aspect ratio is given otherwise. For a colour image, the single image consists of three separate frames, depending on the model used. So, a 480p image



FIGURE 2.1: Images as seen by the computer. Numbers represent pixel values in images (Deshpande 2016).

consists of three 848×480 pixel arrays. The most common colour models are the RGB model and the HSB model. RGB model divides the colour image into shades of Red, Green, and Blue. These three images are stacked to make a colour image, as the combination of red, green, and blue lights can make any colour. Similarly, the HSB model works where the image consists of hue, saturation, and brightness arrays. Despite their abstract nature to human interpretation, these numerical representations ultimately form the foundational input for the computer's image understanding and classification processes.

2.2 Object Detection

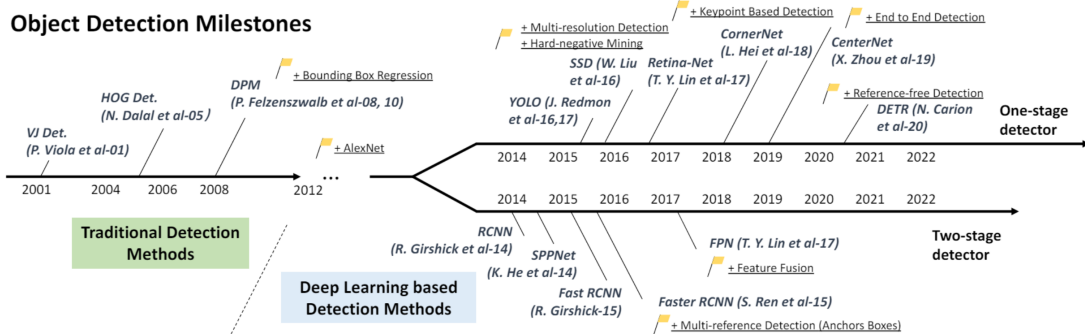


FIGURE 2.2: The image shows a rough timeline of the innovation of object detection algorithms (Zou et al. 2023).

A major research topic in computer vision is focused on detecting objects within images. Most of the time, computer algorithms are focused on detecting animals, humans, road signs, or

vehicles. The prime objective of object detection incorporates devising computational models and methods that address the fundamental question for computer vision applications, such as identifying objects and their spatial placement in the images. To gauge the effectiveness of these algorithms, accuracy in classifying different objects and localisation precision alongside speed stands as the primary yardstick for evaluating object detection methodologies.

Object detection serves as the foundation for numerous other computer vision tasks such as instance segmentation (He et al. 2017), image captioning (Karpathy and Fei-Fei 2015), and object tracking (Kang et al. 2017). Deep learning methodologies (LeCun et al. 2015) have significantly propelled advancements in object detection, breaking barriers and pushing object detection into a highly sought-after research field and garnering unprecedented attention. One of the primary reasons is that object detection has a broad engineering application and has multiple uses in real-world scenarios. Object detection is utilised in autonomous driving, robot vision, and video surveillance. In the past 25 years, object detection has experienced exponential growth in publications.

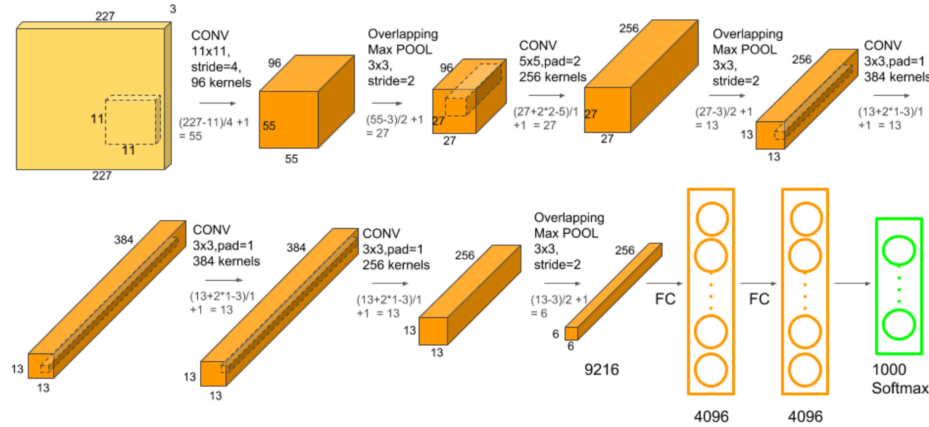


FIGURE 2.3: An illustration of the Alex-net deep neural network structure. Alex-Net comprises 5 convolution layers, 3 max-pooling layers, 2 normalised layers, 2 fully connected layers, and 1 SoftMax layer. Each convolution layer involves a convolution filter and the ReLU non-linear activation function. Max-pooling is applied in the pooling layers, and the input size is maintained as 224x224x3, though padding results in an effective size of 227x227x3. With over 60 million parameters, Alex-Net is a substantial neural network architecture (Al Hassan 2018).

Gathering background knowledge in the historical development of object detection algorithms is essential. I have divided the history of object detection research into two parts. I will briefly discuss both. The first part is the pre Alex-net (Krizhevsky et al. 2012) era. The second part is the modern object detection algorithm era after Alex-net. In the early 2000s, the groundbreaking Viola Jones detector significantly improved real-time human face detection. The detector outpaced other algorithms by tens and even hundreds of times in speed while maintaining compatible detection accuracy. The detector employs the sliding window approach by exploring all potential locations and scales within an image to identify human faces. Although it seemed simple then, the computational demand was challenging to meet with the available computational resources at that time. The VJ detector improved its speed by integrating three crucial techniques: integral image, feature selection, and detection cascade (Viola and Jones 2001).

After Viola Jones in around 2005, N. Dalal and B. Triggs introduced histogram of oriented gradients (HOGs) feature descriptor (Dalal and Triggs 2005) presented a significant improvement over the scale-invariant feature transform (Lowe 1999), and shape context (Belongie et al. 2002). The HOG algorithm operates on a densely spread grid of cells and applies overlapping local contrast normalisation on blocks. Pedestrian detection primarily motivated this, and its flexibility extends to detecting various object classes. This is achieved by rescaling the input image multiple times while maintaining the size of the detection window. HOG detector was instrumental in laying the foundations for numerous other detectors (Felzenszwalb et al. 2008; Malisiewicz et al. 2011) and has found its way into various applications in diverse fields of computer vision over the years.

Next came the Deformable parts model (DPM) (Felzenszwalb et al. 2008). The DPM won the Pascal Visual Object Classes competition in 2007, 2008 and 2009 in the object detection challenges (Everingham et al. 2010). The authors initially proposed it as an extension of the HOG detector, applying the divide and conquer algorithm. The algorithm decoded the objects into their individual parts; for example, if you are trying to detect a human, you would try to individually detect its face, hands, shoulders, and neck and then ensemble them into an object

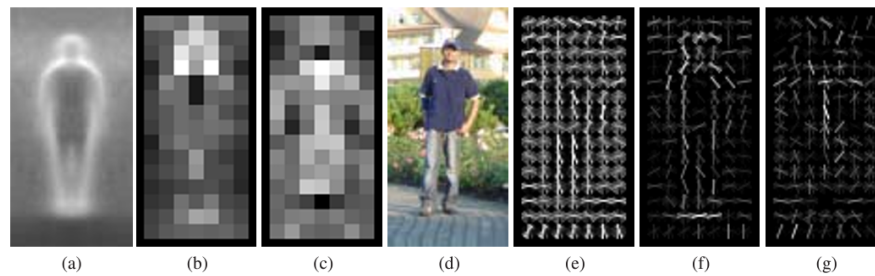


FIGURE 2.4: HOG detectors concentrate mainly on silhouette outlines, particularly on the head, shoulders, and feet. The blocks that exhibit the highest activity are positioned on the image's background just beyond the contour. (a) The gradient image average across the training examples. (b) Each "pixel" showcases the highest positive SVM weight within the block centred on it. (c) Similarly, this applies to the negative SVM weights. (d) An image for testing. (e) The R-HOG descriptor is computed from it. (f, g) The R-HOG descriptor is adjusted according to the positive and negative SVM weights (Dalal and Triggs 2005).

you're trying to identify. The same authors also improved their work by making significantly more contributions and changes to their proposed work in the following years.

2.2.1 Two Stage Detectors

The Large-scale Image-Net challenge in 2012 was instrumental in propelling a resurgence of convolution neural networks, especially led by Alex Krizhevsky (Krizhevsky et al. 2012). This opened a new age of artificial intelligence based on deep neural networks. This innovation spurred unparalleled growth in the future of object detection. Girshick et al. introduced the regions with CNN features in 2014, known as RCNNs (Girshick et al. 2014). From this point onwards, deep learning took over object detection. We categorised deep neural network-based algorithms into two broad categories. (i) two-stage detectors, (ii) one-stage detectors. The two-stage detectors looked at the detection problem as going from course to fine features when it came to detecting objects; alternatively, one-stage detectors looked at detection as a one-step process.

RCNN adopted a simple approach by initiating object proposals generated by selective search (Uijlings et al. 2013). Each potential proposal is resized to a fixed size before being passed

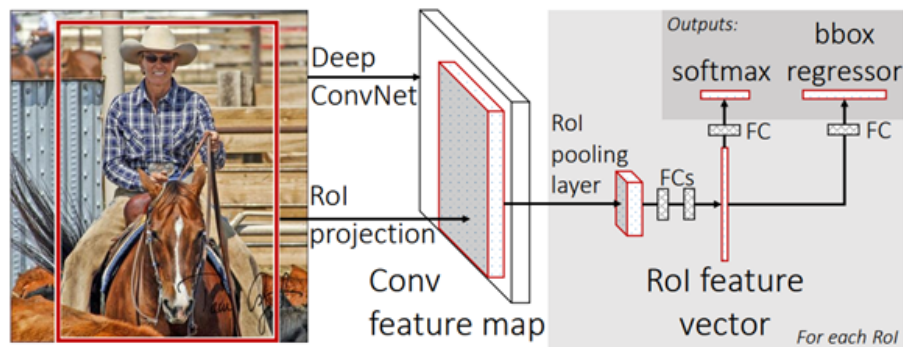


FIGURE 2.5: Figure 2: Fast R-CNN uses a fully convolutional network to process images and regions of interest (ROIs). ROIs are pooled into a fixed-size feature map and transformed into feature vectors by fully connected layers, producing softmax probabilities and bounding-box regression offsets. The architecture is trained end-to-end with a multitask loss.(Girshick 2015)

through a CNN pre-trained on Image-Net dataset (Deng et al. 2009) for feature extraction. After feature extraction, linear support vector machine classifiers determine each region's object presents and categories. I see an increase in the mean average precision from its competitors by a whopping ~28%. One of its major drawbacks was the detection time, which was around 14 seconds per image using a GPU. However, SPP-Net solved the problem the same year RCNN was proposed.

SPP-Net (He et al. 2015) proposed by K. He et al. in 2014 introduced Spatial Pyramid Pooling (SPP) layers, enabling a CNN to generate fixed-length representations irrespective of image or region size without rescaling. SPPNet computer feature maps for the entire image just once and then generated fixed-length representations for various regions; it was able to achieve over 20 times faster detection than RCNN, all the same, maintaining the mAP of ~59.2% in comparison to the accuracy of RCNN of 58.5%. The speed improvement that SPPNet provided had a couple of limitations, such as multistage training and overlooked previous layers, which was solved the following year by Fast RCNNs (F-RCNN) (Girshick 2015).

In 2015, R. Girshick proposed Fast RCNN, a refinement of R-CNN and SPPNet. Fast RCNN allowed simultaneous training of a detector and bounding box regressor in the same network setup. This new algorithm trained on the VOC07 dataset raised the mean average precision from 58.5% (RCNN) to 70%, with detection speeds 200 times faster than RCNN. Although it

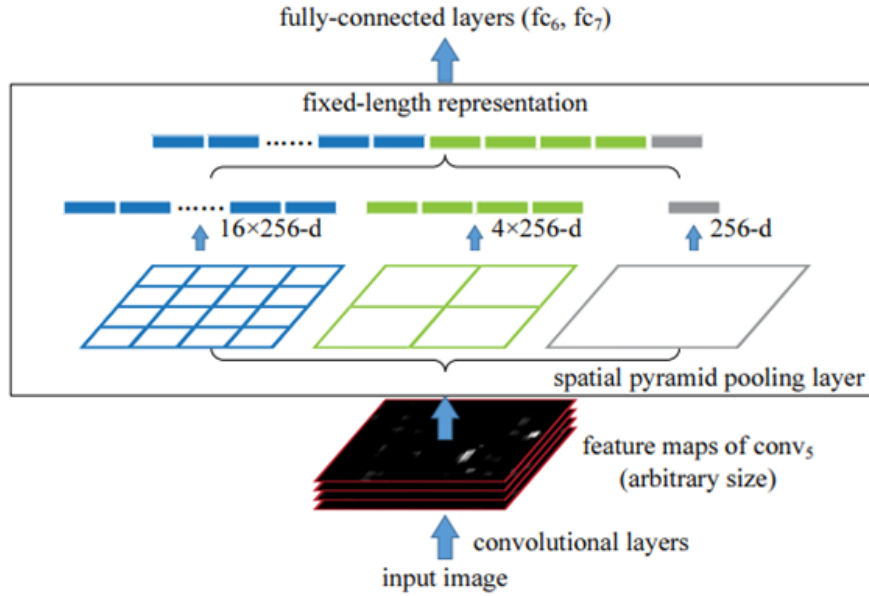


FIGURE 2.6: A network structure with a spatial pyramid pooling layer ((He et al. 2015)).

was a significant improvement from previous works, the detection speeds remained hampered by proposal detections, which led to the birth of Faster R-CNN (Ren et al. 2015). Faster R-CNN, proposed by Ren shortly after Fast RCNN, became the first near-real-time deep learning detector. It achieved a mean average precision of 42.7% considering its intersection over union of above 50% on the COCO dataset and achieving a mean average precision of 73.2%. Trained on VOC07dataset, ZF-Net (Zeiler and Fergus 2014) achieved 17fps.

One of the biggest blockades regarding real-time detection was the region proposals for object detection. Faster R-CNN solved this problem by introducing a region proposal network. From RCNN to Faster RCNN, many components of object detection systems were combined as an end-to-end deep learning framework. All these networks were able to solve the problem of speed. However, there were some issues in terms of different detection stages. One of the biggest one was redundancy always persisted in subsequent detection stages, leading to subsequent enhancements like RFCN (Dai et al. 2016) and Light head RCNN (Li et al. 2017).

FPN (Lin et al. 2017a) revolutionised object detection by leveraging a top-down architecture with lateral connections to form high-level semantics at all scales. FPN's integration into

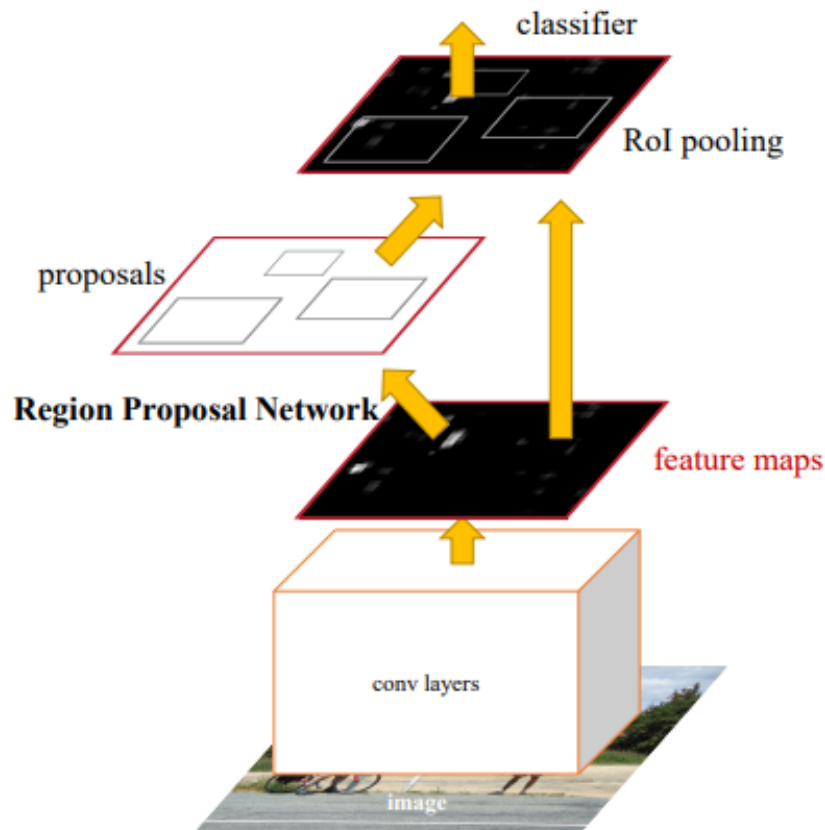


FIGURE 2.7: Faster R-CNN is a unified network for object detection, where the Region Proposal Network (RPN) acts as the 'attention' mechanism in this integrated framework(Ren et al. 2015).

Faster R-CNN systems achieved state-of-the-art single-model detection results on the COCO dataset, becoming a fundamental component in many recent detectors.

2.2.2 One Stage Detectors

In contrast to the two-stage detectors, where. In the first stage. The detectors detected images in a broader sense, a refinement process focused on differentiating these objects and helping these algorithms achieve higher precision. Although these algorithms were exact, they were not used in engineering applications as they were slow.

One-stage detectors were designed to apply these algorithms in the real world, where these object detectors performed object detection in a single step. These algorithms were favoured

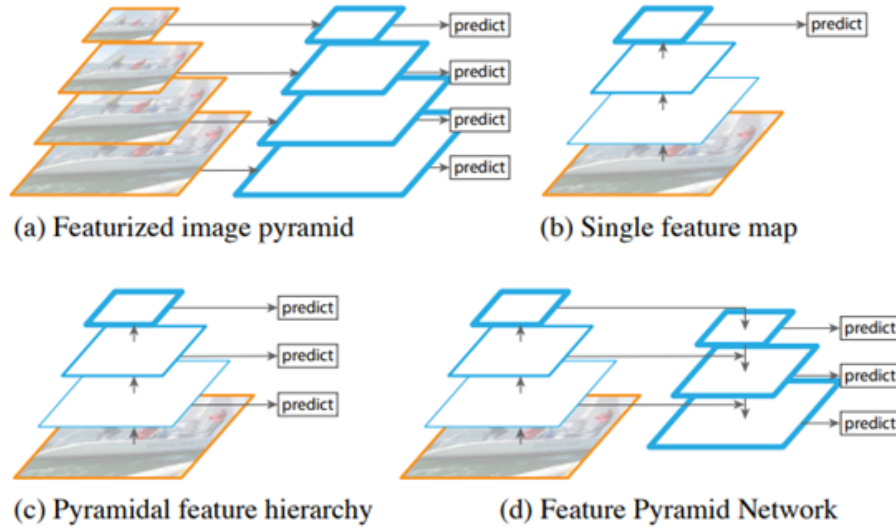


FIGURE 2.8: Illustrates different approaches for feature pyramids in object detection: (a) The traditional method uses an image pyramid, which independently computes features on each scale, resulting in slow processing. (b) Recent systems use only single-scale features for faster detection. (c) An alternative approach reuses the pyramidal feature hierarchy computed by a ConvNet as if it were a featurized image pyramid. (d) The proposed Feature Pyramid Network (FPN) is depicted as a fast and accurate solution, combining the speed of (b) and (c) with improved accuracy. The figure uses blue outlines to represent feature maps, with thicker outlines indicating semantically stronger features. (Lin et al. 2017a)

for deployment in the real-world environment due to the swiftness of their detection in real-time. However, compared to the two-stage detectors, they were not as accurate and struggled to detect tiny objects initially.

In 2015, Redmond Joseph proposed (YOLO) You Only Look Once ushered a new age of one-stage object detectors in deep learning algorithms (Redmon et al. 2016). It achieved the mean average precision of 52.7% on VOC07. The fastest version of YOLO works at 155 frames per second. At the same time, the more precise version works at 45 frames per second, with a mean average precision of 63.4%. YOLO was very different from the two-stage detectors because it used one single neural network for the entire image. The network divided the image into regions and concurrently predicted bounding boxes and the probabilities of finding objects in each region. This enabled the algorithm to perform at very

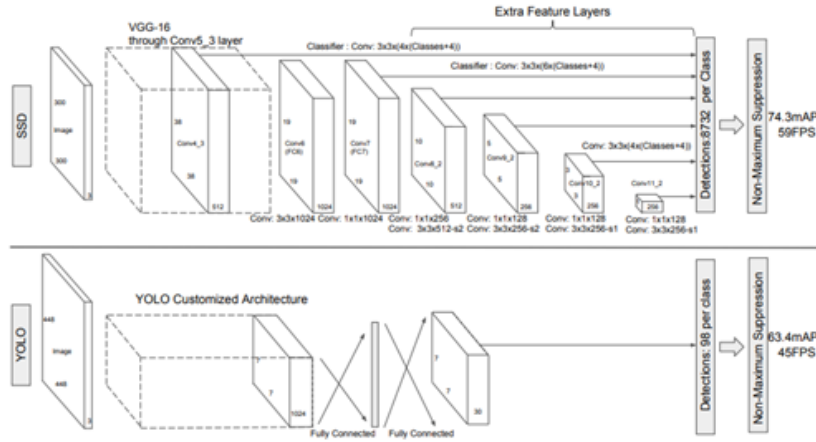


FIGURE 2.9: The YOLO (You Only Look Once) Detection System involves a straightforward process: (a) Resize the input image to 448×448 . (b) Run a single convolutional network on the image. (c) Threshold the resulting detections based on the model's confidence (Redmon et al. 2016). Meanwhile, the SSD (Single Shot MultiBox Detector) model extends a base network by adding additional feature layers at the end. These layers predict the offsets to default boxes with various scales, aspect ratios, and associated confidences (Liu et al. 2016c).

high speeds; however, this approach affected its localisation accuracy in comparison to the two-stage detectors, especially when keeping smaller objects in mind. In the other versions of (Jiang et al. 2022) and SSD algorithm, researchers were working on solving this problem. One of the latest versions, YOLOv7 (Wang et al. 2023), outperforms many existing object detectors in speed and accuracy. It has introduced the concepts of optimised structures, such as dynamic label assignment and model structure reparameterisation.

Introduced in 2015 by Lou et al., the Single-Shot MultiBox Detector (Liu et al. 2016c) brought significant advancements by incorporating multi-reference and multi-resolution detection techniques. It has enhanced the accuracy of one-stage detectors, primarily focusing on detecting small objects. SSD achieved the mean average precision of 46.5% on the Coco dataset, with another version achieving detection rates of 59 frames per second. SSD was one of the first object detectors to detect the object at various scales rather than just working within the top layers.

It addressed the huge difference in accuracy between the one-stage and two-stage detectors and formed the basis of Retina Net, which was introduced in 2017 (Lin et al. 2017b). Researchers proposed the focal loss function, which aimed to diminish the foreground and background class imbalance encountered by these different algorithms during training. This Cross-entropy loss function focused on the challenging and misclassified examples during training, enabling the one-stage detectors to achieve accuracy at par with the two-stage detectors while also preserving high detection speeds by achieving a mean average precision of 59.1% on the Coco dataset.

CornerNet, produced by (Law and Deng 2018), was the next breakthrough in scientific rigour. The algorithm departed from the conventional approach of using anchor boxes for classification followed by regression for object detection. CornerNet looked at object detection as a critical point prediction challenge and focused on predicting bounding box corners. It addressed the drawbacks of category imbalance and manual hyperparameter settings and for long convergence time by decoupling and regrouping corner points with additional embedded information. CornerNet was the most superior one-stage detector of its time; it achieved a mean average precision of 57.8% on the Coco dataset (Lin et al. 2014).

In 2019, Duan et al. proposed the CenterNet (Zhou et al. 2019) using the essential point-based detection technique. CentreNet streamlined post-processing steps like grouping key point assignments and non-maximum suppression. CenterNet considered an object representation as a single point that was its centre and then regressed all the attributes of the object based on the reference point. CenterNet integrated various tasks such as 3D object detection, human pose estimation, optical learning, and estimation into a unified framework. It achieved 61.1% mean average precision on the Coco dataset.

In 2020, N. Carion et al. introduced (Detection Transformer) DETR (Carion et al. 2020a), marking a shift in object detection by leveraging Transformers in deep learning for computer vision. DETR Looked at object detection as a prediction problem by relying on Transformers to replace traditional convolutional operators and enable attention-based competition for a scale-perceptive field. This made anchor boxes or points useless and ushered in a new phase in object detection. Later, deformable DETR was also proposed to overcome the limitations

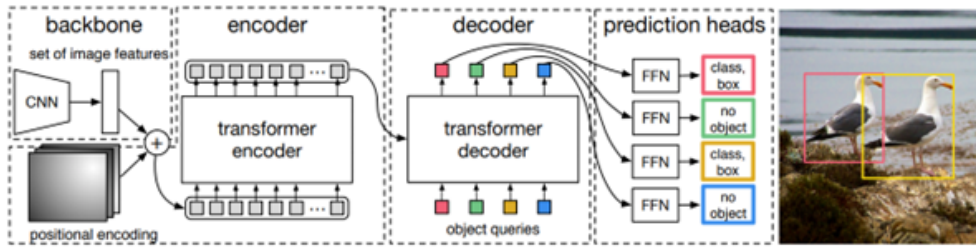


FIGURE 2.10: DETR employs a standard CNN backbone to acquire a 2D representation of an input image. This representation and positional encoding are fed into a transformer encoder. The transformer decoder takes object queries (learned positional embeddings) and attends to the encoder output. Each decoder output embedding is then passed through a shared feedforward network (FFN) to predict either a detection (class and bounding box) or a “no object” class (Zhu et al. 2020c).

of convergence time and to detect small objects; DETR achieved a mean average precision of 71.9% on the Coco dataset. After DETR, many variations of the DETR algorithm have been proposed with improved results in object detection.

2.2.3 Basic Building Blocks and Techniques for Object Detection

Most techniques involved in developing object detection algorithms first focused on three essential factors. Firstly, they were focused on improving the localisation of object detection. Secondly, they were focused on improving the speed of detection, and the last part was making sure that the classification was correct. In this part of my thesis, I will discuss some of these algorithms' techniques and building blocks and their development in the past few years.

One of the main problems in object detection is that when an image or video is passed through the computer, there are many different objects with various sizes and labels. Another issue is that the image being passed through is not the same size. In the beginning, researchers were using the sliding window approach, where they would slide a window of a particular object, as the object is just a combination of pixels and different values of those pixels in the window. Now, the job of the detection algorithm is to say that the object was present in the image at a particular location to match the pixel values by calculating the difference between the object pixel values and the pixel values current in the image and finding the difference. Where the

difference would be minimal, the algorithm would say that there is. This probability that the object is present in this location of the image.

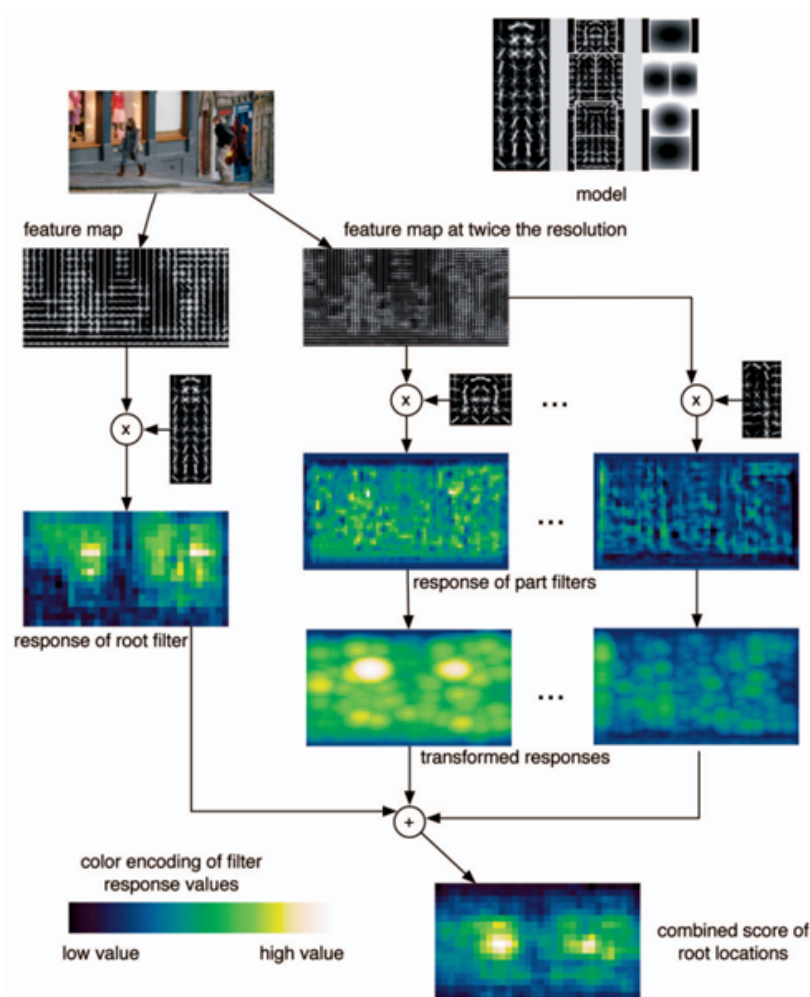


FIGURE 2.11: The matching process at a single scale involves computing responses from root and part filters at different resolutions in the feature pyramid. These responses are transformed and combined to produce a final score for each root location. The example illustrates responses and transforms responses for the "head" and "right shoulder" parts, emphasising the discriminative nature of the "head" filter. The combined scores reveal two viable hypotheses for the object at this scale. (Felzenszwalb et al. 2009)

After the development of HOG & DPM algorithms, R. Girshick et al. proposed a 'Discriminatively Trained Part-Based Model'. It detected more complex objects with different complex appearances, and what it did was that he trained multiple detectors for objects of various

aspects, sizes, and ratios. He also used non-maximal suppression post-processing to get a better result (Felzenszwalb et al. 2009).

Next came another technique of object proposal regions; this algorithm would work on a group of reference boxes placed strategically on the image likely to contain objects. The main objective of this technique was to speed up the object detection process and avoid the sliding window technique, as sliding window would take more time to process (Hosang et al. 2015).

Around 2015, a learning-based top-down approach emerged after the popularity of deep convolution neural networks to address visual recognition tasks. It demonstrated better results in addressing object detection problems. However, in future years after the publication of SSD and YOLO, one-stage detectors, object proposal gradually receded from the forefront of research (Erhan et al. 2014).

After that, because of the growing computational power of GPUs, most of the research focused on multi-scale detection methods. Regression had taken the forefront of research to tackle the multi-scale challenge of object detection. Object detection algorithms directly predict bounding boxes on images based on deep features extracted using convolution neural networks (Yang et al. 2007). Around 2018-2019, a new idea arose where researchers started treating object direction problems as critical point detection problem. Two different methods primarily took hold in the research community, with one method revolving around a group-based approach by identifying key points such as corners or centres representative points of objects and then conducting object-wise grouping (Yang et al. 2019). The second method was a group-free approach, which treated objects as one or multiple points and then performed regression based on size and ratio, considering the reference points (Adams and Bischof 1994).

These days, multi-reference detection is the predominant approach for scale detection, widely used in object detection (Lin et al. 2017a; Liu et al. 2016c; Tian et al. 2019). This technique defines a set of references, usually called anchor box points across every location within an image; after that, bounding boxes are predicted based on these designated references. Another method coupled with it is multi-resolution detection, which involves detecting objects of

different scales and sizes across different layers of the convolution network. Both strategies of multi-reference and multi-resolution have significantly improved the efficiency and accuracy of object detection algorithms.

The following few parts of this writing are instrumental in driving me to research object detection algorithms. When we look at different objects, our brain automatically associates different objects with each other. Our brain then relies on these environments and associations to enhance visual perception and cognition. In my first few experiments that I performed back in 2019 when I was about to explore this field, I saw that the success of object detection algorithms heavily relies on object association, especially when deploying these algorithms in engineering systems in real-world environments (Divvala et al. 2009).

Now, when it comes to object detection, we use the global context, where the scene configuration adds information to help with object detection. Initially, this was incorporated into object detection by providing a statistical summary of the scenic elements (Divvala et al. 2009). However, global context is also incorporated these days by leveraging deep convolution neural networks, deformable convolution, or pulling operations by acquiring receptive field. Sometimes even more significant than the input image (Liu and Huang 2018). Attention-based algorithms and Transformers have also achieved full image receptive fields, improving object detection accuracy (Zhu et al. 2020c). The other method was detection using local context by incorporating surrounding visual information, which is beneficial for object detection.

A recently discovered new domain has emerged as the context interactive domain, where researchers explored the interrelationship between objects and their connectedness that affected the correct classification of objects. For example, we are trying to estimate the age of a person riding a slide in a playground. Let us say that a person's age is around 70 years old. Now, this is unique in the sense that most of the images out there in datasets are of children riding slides. Suppose the algorithm is trained on such data where there is no instance of an older adult riding a slide. In that case, there is a high probability that the algorithm will not correctly estimate the person's age because of the context and environment. This was one of the motivations for this work, as we wanted to see the relationship between age estimation, gender classification, and clothing people wear in different environments and

contexts. He wanted to see the effects of clothing on age estimation. Also, the relationship between different types of clothes and the impact of their detection and classification on other objects (Gupta et al. 2015; Hu et al. 2018; Liu et al. 2018).

I will briefly touch on some other important aspects of object detection. One of the essential concepts is hard negative mining, which involves identifying and prioritising difficult and misclassified negative samples during training to improve the accuracy of the classifier. Imagine a scene with a model that must identify various objects on streets for autonomous vehicle driving systems. The model needs to differentiate between additional cars, bicycles, and other objects to ensure safe driving of the vehicle in real-world environments where you've got partial object occlusion and similar patterns, such as shadows of pedestrians. In these environments, it is hard to differentiate different objects. These challenging negative samples are identified in hard negative mining, and consecutive training attrition is given more priority. The model learns from these negatively identified samples to improve its ability to distinguish between actual objects and challenging background elements, so these misclassified objects adjust the learning process during training. Hard negative mining makes object detection models more robust and accurate, especially in the real-world environment (Hosang et al. 2015).

Another concept is bootstrapping during training. In this method, the algorithm is repeatedly trained using subsets of the primary dataset, sometimes by adding new samples or by removing old ones, to gauge the model's accuracy every time. This helps avoid overfitting and improves the generalisation of the model. For example, consider a dataset you created to detect different objects on the table. Researchers train SSD, YOLO, or Faster RCNN on any other such algorithm or any such other algorithm on the dataset.

After initial training, you first create random subsets of the primary dataset and replace your training data with one of the sample subsets from the primary dataset. He trained the algorithm again on that dataset and checked its accuracy against the validation set. This creates a new version of your model with its precision and accuracy. You repeatedly try this method but the different datasets to see whether the model's accuracy improves. After multiple iterations

of this type of training, you can select one of the best-performing models that demonstrates generalisation and performance on unseen data, and it can be chosen for deployment.

One of the essential concepts in object detection is the loss function, which is defined as the measure of the dissimilarity between predicted and actual values, and it also serves as an optimisation objective during training. What we are trying to do during training is to minimise our loss as much as possible for both classification and localisation.

Different loss functions are used in object detection, and researchers have proposed other loss functions over time. Cross-entropy loss is one of the main loss functions used for classification, and it is used to demonstrate one of the primary examples for classifying whether the animal is a dog or a cat. So, let's say we pass an image to the model, and it is a cat; however, the model classifies it as a dog. This means a very high classification loss, and the model is not trained correctly. The other loss functions that we use are known as localisation loss. The most common ones are L1, L2, and smooth L1 losses. The model predicts where the animal is in the image and draws the bounding box with X&Y coordinates and the width and height of the bounding box. The box should be drawn around the animal or the object being detected. If the predicted bounding box significantly deviates from the ground truth bounding box, it means that there is a high localisation loss present in the model. For instance, if an object's predicted bounding box is slightly misaligned from its actual position, bounding box regression will adjust these coordinates to encapsulate the object better.

One of the last steps involved is known as non-maximal suppression. It is a post-processing technique to filter out redundant bounding boxes predicted by the algorithm by selecting the most confident ones. For example, an object detector might have predicted several bounding boxes for an object in an image. NMS will retain only the ones with the highest confidence scores, while discarding the rest. In most algorithms, the researchers can choose the confidence score of the detection, and depending on the application, it can be 50% or 95%. That means that all the confidence scores less than 50% or 95% depending upon the application are discarded, whereas any scores above the set scale are kept. Some learning-based NMS techniques have recently been introduced in object detection (Desai et al. 2011).

Researchers are also developing NMS-free detectors to realise the goal of end-to-end object detection (Zhu et al. 2020c).

2.2.4 Speeding up of Detection

One of the primary goals of science and technology is to make life easy for us humans and to assist us in this world. This is one of the reasons a lot of research has been done on object detection algorithms. However, one of the most significant issues has been the real-world deployment of these algorithms. The reason is computers' limited technology and computational power, which caused a significant bottleneck in object detection speeds for several years. In the next section, I will briefly touch on the different processes and improvements in object detection algorithms that enabled these algorithms to achieve real-time processing speed. Real-time processing is defined as around 25 to 30 frames per second. If an algorithm detects and classifies about 25 to 30 frames per second, it is said to have achieved real-time object detection speed.

Researchers improved the speed thousands of times by targeting various parts of the object detection algorithms. One of the most computationally time-consuming processes in the detection predominantly arises from feature extraction. Researchers primarily focused on shortening the time to extract features from images by removing a feature map from the entire image, thus significantly improving computational efficiency (Girshick 2015). Another method that has been widely used is to make convolution networks more efficient by removing all the unnecessary network structures and by quantifying activations or weights. This is done by removing a few weights after every training is completed, and recently, network quantification has focused on binarisation by converting floating point operations to logical ones (Han et al. 2015). Another way the speeding up of these algorithms has been done by designing lightweight networks (He and Sun 2015) encompassing various approaches such as factorising convolutions (Zhang et al. 2015), grouping convolutions incorporating depth-wise separable convolutions (Chollet 2017), incorporating bottleneck design and employing NAS techniques (Chen et al. 2019).

After the publication of attention is all you need (Vaswani et al. 2017). The use of Transformers overtook the field of artificial intelligence. Initially, Transformers were combined with CNNs for object detection (Zhu et al. 2020c). This laid the groundwork for using Transformers for object detection purposes. The performance of DETR based on ResNet50 significantly outperformed many modern object detectors and matched the performance of Faster-RCNNs. The main components of DETR were a CNN-based backbone followed by an encoder-decoder transform that processed the one-dimension feature map obtained from the convolution neural network and a simple feedforward network for decoding the fixed-level embeddings into the bounding box ordinates. The feedforward neural network predicts normalised centre coordinates, height width, and category labels via soft-max layered linear function (Yang et al. 2007).

2.3 Face Detection and Alignment Techniques

This subsection will discuss face detection algorithms, alignment techniques, and our implementation methodology. (Feng et al. 2021b) categorised the deep learning face detectors into three main groups. Multi-stage face detection frameworks: these were initially inspired by cascade classifiers in traditional spaces; this was the main point of initial exploration of applying deep learning techniques to face detection. After that, the two-stage face-detecting algorithms, in which the first stage involved generating proposals and the second stage confirmed whether these proposals had a face or not. Lastly, one-stage face detection frameworks integrated feature extraction and proposal generation into a unified network.

2.3.1 Multi-Stage Face Detectors

At the beginning of using deep learning for face detection, most deep learning algorithms were based on detectors designed with multiple stages, also known as a cascade structure. This is why most of them are known as cascaded face detector algorithms. In these algorithms, the face undergoes feature extraction and refinement one after the other, before being put into a classifier and then a regressor. The primary purpose is to discard the sliding window

technique, enhancing overall efficiency. Li et. al (Li et al. 2015) demonstrated an early approach to this type of network. They introduced cascaded CNN, a face detector based on a convolution neural network, which achieved reasonable results on the FDDB dataset (Jain and Learned-Miller 2010).

One of the other notable implementations of cascaded CNNs for face detection is widely recognised as MTCNN (Zhang et al. 2016). It comprises three subnetworks: P-Net for generating candidate facial windows and layers capable of processing images of different sizes. R-Net is for discarding false candidates and refining the remaining ones using a multitasking approach; O-Net produces the final output, including face-bounding boxes and landmarks. MTCNN achieved significant success by demonstrating state-of-the-art performance on datasets like WiderFace (Zhang et al. 2016).

2.3.2 Two Stage - Face Detectors

In the context of two-stage network architectures inspired by object detection, a crucial component is the region proposal network (RPN) (Ren et al. 2015). However, RPN output features cannot address this problem for tiny faces. CMS-RCNN (Zhu et al. 2017) was proposed, incorporating a contextual multi-scale design in RPN in the final detection stage.

Face RCNN (Wang et al. 2017) based on the Faster-RCNN architecture incorporates Centre loss, online hard negative mining, and a multi-scale training strategy. F-DR Net (Liu et al. 2016a) uses multi-scale training and testing and a non-maximum suppression strategy based on voting faster R CNN. It also features a lightweight design. Face R-FCN incorporates position-sensitive average pooling assigned by varying weights to different face parts based on R-FCN (Dai et al. 2016).

2.3.3 One stage - Face detectors:

In contrast to two-stage detectors, one-stage face detectors can carry out feature extraction and proposal generation and place detection in a single unified convolution neural network.

These are deployable in real-world environments, as their runtime efficiency is far better than two-stage detection. Dense anchors (Najibi et al. 2017) replace face-detectors region proposals in one stage. HR, (Hu and Ramanan 2017) introduced by Hu et al., is among the pioneering methods to conduct anchor-based face detection within a unified convolutional neural network. The ResNet-101 (He et al. 2016) serves as the backbone for HR. Contrary to HR, SSH can detect faces at varying scales of different feature levels. The module has been widely adopted in subsequent works (Deng et al. 2020; Li et al. 2019; Tang et al. 2018).

One-stage face detectors like S3FD (Zhang et al. 2017a) based on the extended VGG16 backbone incorporated additional convolutional layers to detect various faces at different scales. Pyramid box also uses the same backbone, enhancing the features with FPN (Lin et al. 2017a). It also incorporates a deeper neural net with an improved SSH module inspired by Inception Resnet and DSSD (Fu et al. 2017; Szegedy et al. 2017). DSSD contained the feature-enhanced module for refined detection. It also inherited the S3FD backbone. SRN (Chi et al. 2019) employed a Dual shot network, refining low-level features with a 2-step classification and high-level features with two-step regression using Cascade RCNN (Cai and Vasconcelos 2018).

Retina-face, another algorithm using anchor-based methods such as FPN adding an extra pyramid layer and employing deformable convolution neural networks (Dai et al. 2017). The DCN module was used for lateral connections and context modules, incorporating a 3D mesh with around 1000 points. It also had a 5 landmark mask and a bounding box using cascade regression for improved localisation.

After introducing Corner-Net (Law and Deng 2018), research focused on exploring anchor-free object detection models and designs. CSP models a face with a bounding box and centre point, leveraging the multi-scale features from modified Resnet-50 (He et al. 2016). Application of this technique can capture rich local and global features and information around the face. Progress-Face (Zhu et al. 2020b) improves the one-stage networks by adding an anchor-free module, addressing the imbalance of positive and negative samples for small faces.

Recently, one-stage networks have become more prevalent in face detection for several reasons. One-stage detectors have gained popularity in face detection due to their runtime independence from the number of faces, computational efficiency for achieving near-scale invariance, and adaptability to face detection's relatively less complex nature than general object detection. Face detection is a task of relatively low complexity compared to available object detection, which has allowed researchers to swiftly adapt object detection to face detection by considering unique patterns of faces.

2.4 Overview of Datasets in Object Detection

In this section, I will start by discussing generic object detection datasets and some collection methodologies for creating these datasets. This will be followed by a brief discussion of age and gender datasets that are available and developed by different researchers. Toward the end, I will discuss attire, which is usually considered fashion detection in computer vision.

In the age of deep learning for object detection, developing sophisticated algorithms heavily depends on creating unbiased, extensive, comprehensive datasets. Over the last few years, many different object detection datasets have been introduced, including Pascal-VOC challenges, the ImageNet large-scale visual recognition challenge (Chao et al. 2009), Coco detection challenge (Lin et al. 2014) Open image dataset challenge (Kuznetsova et al. 2020), and various others.

The CIFAR-10 dataset (Krizhevsky and Hinton 2009), a subset of tiny images, includes 60,000 32x32 colour images categorised into 10 distinct classes: aeroplane, automobile (excluding trucks and pickups), bird, cat, deer, dog, frog, horse, ship, and truck (excluding pickups). Each class comprises 6,000 images, split into 5,000 for training and 1,000 for testing. The criteria for labelling this dataset were as follows: the annotators asked questions such as, what is in this picture? Each image should predominantly feature a single instance of the object. The object could have been partially obscured if its identity was discernible to an annotator (Krizhevsky and Hinton 2009).

One of the pioneering datasets in every detection was the Pascal visual object classes, also known as VOC challenge (Everingham et al. 2010), carried out from 2005 to 2012. Research in object detection was in its early stages in the computer vision community. This competition challenged the computer vision research community and motivated researchers to research this field.

The PASCAL VOC challenge provided two sets of data. One was known as the VOC07, which constituted 5000 training images and around 12,000 annotated objects, whereas the VOC12, the other, consisted of 11,000 training images and about 27,000 annotated objects. The annotators annotated around 20 common object classes, such as a person, cat, bicycle, sofa, etc. The scenario in these images was also very similar to everyday, common images. There were two main challenges in this competition: one was classification, which was to determine the presence or absence of the 20 classes in test images, and the second one was detection, which was predicting the bounding box coordinates and labels for each of the objects amongst the 20 classes within the test image.

Another significant challenge was the image net large-scale visual recognition challenge, also known as ILSVRC (Krizhevsky et al. 2012), which was very important in propagating the research in generic object detection. It started in 2010 and ended in 2017. It was featured as an object detection challenge using internet images with around 200 different classes of visual objects compared to VOC, a significantly large dataset with many more classes. It had over 40 million annotated images and annotated bounding boxes available in almost at least 1,000,000 images. The modern ImageNet dataset now contains around 20,000 categories, each containing several hundred images.

MS-COCO dataset (Lin et al. 2014), also known as Microsoft Common Objects in context dataset, is one of the most comprehensive and widely used dataset. Since 2015, an annual competition has been performed on the dataset, challenging computer vision researchers worldwide. Compared to ILSVRC, it has fewer object categories but a more significant number of object instances. For example, MS-COCO 17 comprises around 164,000 images and 897,000 annotated objects distributed across 80 classes.

MS-COCO-Dataset is unique in its annotation methodology as the annotators meticulously labelled every different object detection using pre-instance segmentation that significantly enhance the precision of localisation of these objects. Moreover, in this dataset, there are a lot of small objects in vision in comparison as the image size is less than 1% of the image, and they're also very densely clustered together, which poses a more complex detection challenge for computer vision researchers. MS-COCO dataset has emerged as a standard in the object detection domain, pivotal in testing state-of-the-art object detection algorithms and models.

Around 2018, the Open-Image Detection challenge (Krasin et al. 2017) was introduced, which marked a significant leap in object detection datasets. There were two main challenges in this dataset. The first was standard object detection, and the second was visual relationship detection, focusing on identifying objects with specific relations. The dataset had around 1,910,000 images and about 15,440,000 annotated bounding boxes across a range of about 600 object categories.

2.4.1 Face Detection Datasets and related work

FDDB, the face detection dataset and benchmark, emerged around 2010. It was instrumental in benchmarking different face detectors (Jain and Learned-Miller 2010). The images were collected from the Yahoo news website, duplicates were filtered out, and around 2845 unique images were selected. The criteria for choosing the faces were that the minimum height and width of the face should be more than 20 pixels, both eyes should be visible, and the angle between the nose and the camera should exceed 90° . Also, it was left out if a human could not recognise a face. After applying these criteria, annotations of around 5171 faces were left, annotated by elliptical regions covering the forehead to the chin vertically and from left to right on the cheeks horizontally. FDDB propelled unconstrained face detection and addressed expression, pose, scale, and occlusion challenges. This dataset had a strong bias; it was heavily leaning towards celebrity faces, owing to its collection from news sources.

In 2012, detecting faces in the wild dataset was published (Zhu and Ramanan 2012). This dataset was constructed by randomly extracting images from Flickr, focusing on images that

contained at least one sizeable face. The dataset had around 205 images, from which 468 faces were meticulously annotated. The faces were labelled with a bounding box with six specific landmarks.

This was followed by the Pascal Face dataset, which was driven by selecting 851 images from the Pascal VOC test set (Yan et al. 2014). It contains around, 1335 annotated faces. These datasets were introduced by their authors as a benchmark for face detection algorithms. However, because of the small scale of these datasets, they could not train deep learning algorithms to detect faces in the wild.

Young introduced the multi-attribute label face, also known as the MALF dataset. This dataset aimed to comprehensively detect the face and evaluate face detection algorithms in natural environments. The researchers used Beethoven Flickr to collect a dataset of 5250 images with around 11,931 labelled faces (Yang et al. 2015). This was far larger than FDDB, AFW (Zhu and Ramanan 2012), and the Pascal Face dataset. It provided an extensive dataset for face detection evaluation.

The annotation in this dataset was done using axis-aligned square boxes, meaning that the objective of the annotators was to enclose the entire face in the bounding box, ensuring that the nose was positioned right in the centre of the box. If somebody looked sideways, the bounding box on that face would have half of it as background. The dataset also included other annotations, such as gender pose or occlusion, which provided richer information for comprehensive face analysis and opened other avenues of research.

The research community saw 2016 as the year the Wider Face dataset emerged as a pivotal face detection benchmark (Yang et al. 2016). It had 32,203 meticulously curated images across 61 different categories. It was carefully scrapped from popular search engines, and this dataset was divided into training validation and testing subsets. It set a new standard for face detection benchmarks. Wider face annotations aimed at diversifying scale, occlusion, and pose variations; it had bounding boxes encompassing facial features that accommodated occluded faces. It had around 50% of the faces, ranging from 10 to 50 pixels in width and height. It had about 393,703 annotated faces, averaging about 12 faces per image. It was also divided

into three categories, easy, medium and hard levels, to provide detailed testing and evaluation of different face detection algorithms. The unconstrained Face Detection Dataset (UFDD) consisted of 6425 images with around 10,897 face annotations. It had various challenging conditions such as rain, snow, haze lens impediments, blurred illumination variations, and distractors (Nada et al. 2018). Another dataset known as the VGGFace-2 dataset had about 3.31 million images featuring 9131 individuals. This created an average of 362.6 images for every subject in the dataset. These images were usually collected from Google image searches. The dataset had a variety of poses, ages, illuminations, and ethnicities and had a range of different individuals from actors, athletes, and politicians (Cao et al. 2018).

2.4.2 Age and Gender Datasets and influential works

Dataset	Year	# Images	Age Labels	In-the-wild	Attire	Gender
FG-NET(Lanitis et al. 2002)	2002	1,002	to-year	No	No	Yes
MORPH(Ricanek and Tesafaye 2006c)	2006	1,724	to-year	No	No	Yes
IFP-Y (Özbulak et al. 2016)	2008	8,000	to-year	No	No	Yes
Gallagher(Gallagher and Chen 2008)	2009	28,231	7 age groups	Yes	No	No
VADANA (Somanath et al. 2011)	2011	2,298	4 age groups	Yes	No	Yes
AdienceFaces(Eidinger et al. 2014a)	2014	26,580	8 age groups	Yes	No	Yes
CACD (Chen et al. 2015b)	2014	163,446	to-year	Yes	No	Yes
IMDb-WIKI(Rothe et al. 2016)	2015	523,051	to-year	Yes	No	Yes
AgeDB(Moschoglou et al. 2017)	2017	16,488	to-year	Yes	No	Yes
IMDb-AGA* (Ours)	2024	8,000	to-year	Yes	Yes	Yes

TABLE 2.1: Age and gender datasets available, compared to our dataset IMDb-AGA.

To create a large-scale age and gender dataset, researchers created a very large dataset focused on celebrities. They compiled around the top 100,000 actors listed on the IMDb websites and systematically collected their data, including date of birth, name, gender and associated images; this dataset was further enriched by the information provided on Wikipedia. They filtered out the images lacking timestamps to increase the accuracy of the dataset. They created a dataset of around 460,723 face images from around 20,284 celebrities. Adding the images from Wikipedia, of approximately 62,328 images, brought the total to about 523,051 face images. To ensure consistency, the researchers created a threshold using photos where the secondary most robust face detection for below the specified threshold they tried to equalise

the age distribution within the dataset (Rothe et al. 2015). Researchers created the IMDB clean dataset using the IMDB dataset, consisting of around 183,886 training images, 45,971 validation images, and 56,086 test images. Researchers used various methodologies to filter out incorrect labelling and ambiguous annotations (Lin et al. 2022).

Another dataset, CACD, the cross-age celebrity dataset, is available, with around 160,000 facial images representing about 2000 individuals. This is also organised into three subsets of training validation and test sets containing 1800, 120, and 80 individuals, respectively (Chen et al. 2015a). Another dataset known as the KAN-face dataset (Georgopoulos et al. 2020) contains around 41,000 and 36 images of about 1045 subjects ranging from 0 to 100 years. This dataset has many images in different orientations, poses, expressions, and lighting. Another dataset is the Morph dataset (Ricanek and Tesafaye 2006a), which consists of 55,134 mug shot images from around 13,617 subjects aged 16 to 77 years. This is not characterised as in the wild dataset. However, this can train deep learning algorithms to detect ages.

Another face detection age estimation dataset known as UTK-face dataset (Zhang et al. 2017d) is very popular amongst the research community. It contains faces of people ranging from ages 0 to 116 years old. The images showcase various pose variations, facial expressions, lighting conditions, occlusions, resolutions, and other attributes. The dataset can be used for multiple applications such as face detection, age estimation, age progression slash regression and landmark localisation. The researchers captured around 20,000 additional facial images in natural settings. Each image had a single face. The researchers also provide a detailed annotation of 68 points called landmarks to aid in detailed facial analysis. It had comprehensive labels of age, gender, and ethnicity associated with each facial image.

The Fair Face dataset contains around 86,744 training images and 10,954 validation images. It also provides comprehensive ground truth attributes, including race, gender, and age, which are classified into nine distinct classes. It encompasses attributes related to race, gender, and age, categorised into nine classes ranging from infancy to elderly individuals (0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+). Additionally, the dataset is notably well-balanced across different racial groups (Kärkkäinen and Joo 2019).

One of the first datasets was the Audience dataset (Eidinger et al. 2014b) consisting of around 26,580 facial images portraying around 2284 individuals across eight distinct groups ranging from infancy zero to 100 years old. It also included binary gender labels and showcased images captured under diverse conditions encompassing varying head poses, lighting, and image quality.

The AgeDB (Moschoglou et al. 2017) comprises 16,488 images featuring diverse public figures like actors, writers, scientists, politicians, and more. The images are annotated with identity, age, and general information. The dataset has been manually annotated. There are around 568 unique subjects, with an average of approximately 29 images per subject. The age range spans from 1 to 101 years old, with an average age of around 50.3 years for every subject.

Recently, (Bhardwaj and Sengar 2024) presented a real-time system using CNNs and the Caffe framework. It integrated Google Firebase for real-time management and OpenCV for image processing. The use of Firebase enhances scalability and real-time capabilities, making the system ideal for applications like surveillance, marketing, and personalised user experiences. Another similar study (Arora et al. 2024) developed a framework using deep learning for object detection, age estimation, and gender prediction. It integrated the Mask R-CNN model for object detection and the Deep Face library for age and gender estimation.

2.4.3 Attire detection Datasets and influential works

Generally, clothing detection can be divided into categories such as clothing modelling (Chen et al. 2006; Ng and Grimsdale 1996), segmentation (Hasan and Hogg 2010; Wang and Ai 2011; Yamaguchi et al. 2013), recognition (Yang and Yu 2011), recommendation systems (Packer et al. 2018; Zhang et al. 2017b) describing people (Chen et al. 2015c) and classification and retrieval (Bossard et al. 2012; Chen et al. 2012; Lao and Jagadeesh 2016; Lin et al. 2017b). Each category has multiple research problems associated with it. We have decided to focus on the challenge of real-time clothing detection and localisation.

Most of the current research in attire detection is directed by the online fashion and retail industry (Lao and Jagadeesh 2016; Liu et al. 2016b; Packer et al. 2018). Recommendation systems and fashion aesthetics are usually the targets of these works. Their datasets are mostly synthetic and are created from web scrapping (Liu et al. 2016b) online retail clothing websites. Most datasets contain only one labelled clothing per image. Researchers have deployed techniques for detecting clothes or retrieving similar clothing. Most researchers have benchmarked the results on their dataset. However, very few people talk about real-time video implementation of their results. We have found that datasets focused on real-world implementations of attire detection are scarce, so we have decided to address this problem.

In the first half of the last two decades, many attire-related studies have attempted to use different methods to either identify people through clothes (Chao et al. 2009) or used other techniques to identify fashion and clothing textures (Chang and Wang 1996; Chao et al. 2009; Freixenet et al. 2002; Zhang et al. 2008). One of the earliest works that addressed clothing detection in real-time video was presented by (Yang and Yu 2011). The researchers presented a complete system of detecting and segmenting clothes in real-time video. Their main contribution was designing a system tailored for clothing analysis and segmentation. The evaluation included different clothing representations by combining colour (Dalal and Triggs 2005), bag of word features (Yang et al. 2007), discrete cosine transform responses, and region growing method (Adams and Bischof 1994; Liu et al. 2016d) (Freixenet et al. 2002). They devised a complex clothing segmentation method that includes candidate region, canny edge detection, Voronoi image selection, region merging, region growth, and background removal. Their work has many limitations, such as clothing segmentation, which requires proper human figure alignment to feature extraction and classification. They report an average recall of 80% in the top eight clothing categories.

Another influential study, Deep Fashion (Hadi Kiapour et al. 2015; Liu et al. 2016d), also presents a dataset of around 800k images. They develop a neural network that jointly trains and predicts landmarks (their form of annotations) and attributes. The deep Fashion dataset contains attributes of clothing landmarks and cross-pose/cross-domain correspondence of clothing pairs. The dataset includes 50 categories and 1000 attributes; they claim that the

landmark detection is better than the bounding box label presented by (Hadi Kiapour et al. 2015). However, they have branched into many aspects of clothing and fashion representation rather than focusing on one. Their neural network outperforms two bounding box datasets (Girshick 2015; Huang et al. 2015).

A new annotated dataset of 120k images was recently created to detect critical apparel key points, Fashion Landmarks (Ge et al. 2019). Fashion Landmarks are 8 key points that can be used to identify and detect clothing. They selected a large subset of the Deep Fashion database (Chao et al. 2009) to create a new dataset of Fashion landmarks (FLD). They compared the detection of fashion landmarks to the detection of human joints to claim that landmark detection is more complicated than human joint detection. They also proposed a new neural network for landmark estimation. They argued that fashion landmark simulation can help people recognise and retrieve clothes. Researchers used a subset of the deep fashion dataset of 10000 images and 50 clothing attributes. The bounding boxes were manually annotated. The top 5 recall rates of the 50 attributes for the bounding box is 53%. The top 20 and retrieval accuracy for the bounding box is 40%.

Studies simulating clothing have been centred extensively on clothing segmentation versus blank setting (Chen et al. 2006; Hasan and Hogg 2010; Ng and Grimsdale 1996; Yamaguchi et al. 2013) Attempts have also been made on shop-clothing image sorting, applying conventional handmade features (e.g. SIFT, HOG) (Chen et al. 2006; Packer et al. 2018; Wang and Ai 2011; Yang and Yu 2011) to online images and lately deep learning-based characteristics (Bossard et al. 2013; Chen et al. 2015c; Zhang et al. 2017b) have also been used. Turned off by the cost of annotations, consequently a deficiency of significant attire attribute annotations by dissimilar sources, cross-domain dressing attribute learning makes up an intriguing job and, for the most part, not studied (Bossard et al. 2013; Zhang et al. 2017b). Most datasets are synthetic and mined from online clothing websites. Therefore, images are mostly labelled according to their attributes. Other data sets like DCSA, ACWS, DDAN, and DARN (Bossard et al. 2013) don't have any bounding box annotations.

CHAPTER 3

Methods

This chapter explains in detail the construction of the IMDB-AGA dataset. It explains the image selection process from the core IMDB-WIKI dataset. One challenge was ensuring randomness to mimic real-world scenarios and avoiding biases in our dataset. We will also discuss challenges and solutions employed in face detection, alignment, and the manual annotation of age, gender, and attire.

The chapter highlights the different aspects of our dataset, IMDB-AGA, which stands apart in its integration of attire annotation alongside age and gender—attributes that significantly enhances the dataset’s use in the wild. This dataset sets a new standard for future datasets targeting real-world deployment.

Through detailed discussions, visual aids, and comparative analyses with existing datasets, this chapter aims to provide a comprehensive overview of our dataset construction methodologies, the rationale behind our choices, and the anticipated impact of our contributions to the field of computer vision.

3.1 Construction of Dataset

To address **RQ1**, we constructed the dataset IMDB-AGA (IMDb-Age-Gender-Attire)¹, which was done by a systematic and iterative process to ensure that the annotations were of the highest quality. The overview of the decision process can be seen in Figure 3.1.

¹We have made the dataset and code available at: <https://github.com/saadyousuf45/Age-Attire-Gender-Detection-and-Classification>

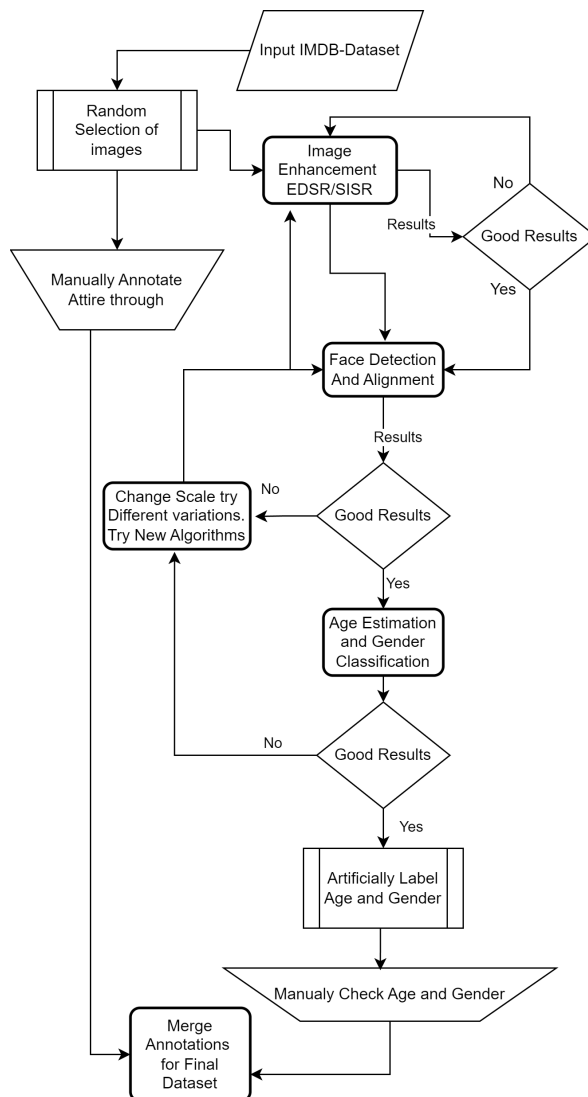


FIGURE 3.1: The flow chart illustrates constructing a multi-label dataset integrating age, gender, and attire annotations using the IMDb-Attire dataset.

IMDb dataset was chosen to form the foundation of our dataset. The dataset comprises around 500,000 images with diverse human facial expressions, attire, and environmental settings, with all the necessary ingredients to build a comprehensive dataset. Firstly, the images were randomly selected; however, to make sure that we had all different types of persons in the images, they were selected at an interval of 50 images. We formed a dataset of around 10,000 images, and the randomness helped us capture different scenarios and subjects, reducing biases. The selected images were manually annotated by annotators, who labelled all items worn by any person in the image. We decided that if a human can identify any attire in

the image, it should be labelled. Another annotator checked the labelling to ensure all the annotations were correct.

To address **RQ3**, We employed image enhancement techniques such as Enhanced Deep Residual Networks (EDSR) and Single Image Super-Resolution (SISR) were applied to improve the quality of the images. Enhanced image quality facilitates better performance in subsequent face detection and alignment tasks. If the enhancement results were satisfactory, the process moved forward; otherwise, further adjustments were made. At every stage, it was ensured that the images were of quality. Adjustments were made if any age or gender results were of poor quality, consisting of changing scale, trying different variations, or looking at different face detection and alignment algorithms.

After successfully detecting and aligning the faces, the next step was to tackle apparent age estimation and gender classification. This involved many experiments with various age and gender estimation algorithms to determine the best-performing model. A small dataset was chosen as a test case, and these algorithms were tested on the dataset to ensure accuracy if the results were unsatisfactory.

Once the best algorithms were identified, they were used to label the age and gender of the detected faces artificially. This provided the annotators with the initial labels and sped up the annotation process while maintaining accuracy. Finally, the annotations of age, gender, and attire were combined from the IMDB-AGA dataset. After that, human annotators manually check the labels. This step is very important to ensure the dataset is accurate and reliable. The annotators reviewed and corrected discrepancies, ensuring the final dataset was accurate and comprehensive.

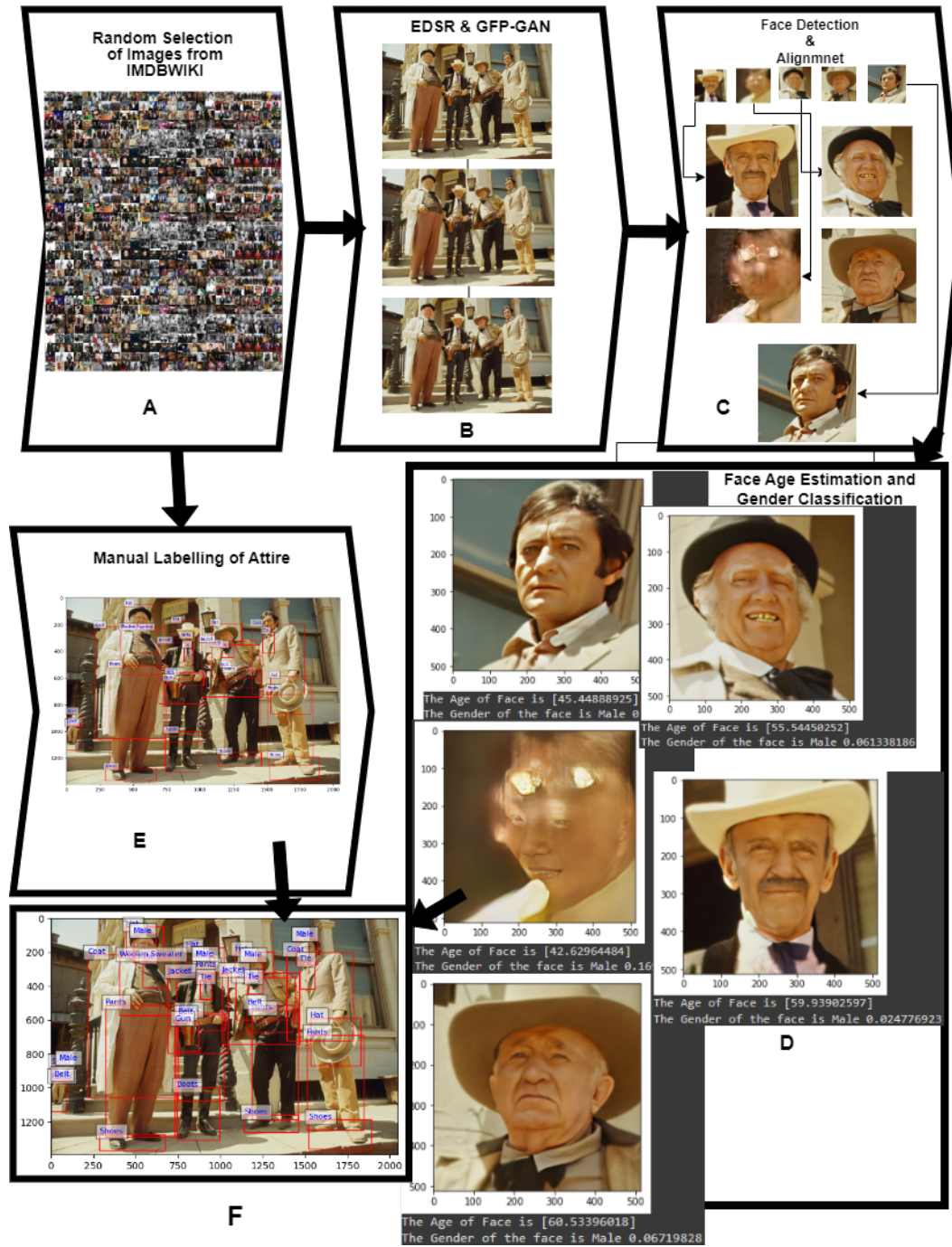


FIGURE 3.2: (A) A dataset of ~8000 is randomly created from the IMDb dataset. (B) Extensive image processing is tested with simultaneously testing face detection and age and gender Estimation experiments. (C) Face detection and Alignment are performed. (D) The age and gender of faces are annotated artificially. Simultaneously, attire is annotated manually. Finally, the labels are combined and manually checked.

The dataset’s construction process started with thoroughly examining existing datasets available at that time. We only looked through the available datasets for age estimation and gender classification. Among the available datasets, such as (Chen et al. 2014; Moschoglou et al. 2017; Ricanek and Tesafaye 2006b; Zhang et al. 2017c; Zhang et al. 2017e), We chose the IMDb-wiki dataset, which consisted of approximately 500,000 images (Rothe et al. 2015). We considered the dataset’s size and performance at the state-of-the-art level. A comparison between available datasets and ours can be seen in Table 2.4.2.

Including attire annotations in IMDb-AGA enhanced its applicability for real-world scenarios compared to previous datasets. Attire information provides additional context that can be of the utmost importance for applications like security surveillance, where identifying a subject’s appearance beyond just age and gender can be vital. Moreover, the in-the-wild nature of IMDb-AGA ensures that models trained on this dataset are robust and generalise well to real-world conditions.

While not the largest, the IMDb-AGA dataset is enough for training advanced models, especially given its comprehensive labelling. IMDb-AGA significantly advances age and gender datasets by integrating attire annotations, providing a more holistic approach to visual recognition tasks. This dataset provides a way to incorporate multi-class labelling and sets a new standard for future dataset construction, promoting the development of more accurate and context-aware models.

3.1.1 Threats to Validity

Reflecting on the dataset collection process, certain requirements would have enhanced the representation and diversity of the final dataset. The absence of stratified sampling raises concerns regarding the equitable representation of various age groups within our dataset. A stratified sampling approach would have addressed potential biases and further contributed to our dataset’s robustness.

Moving forward on this project, the incorporation of stratified sampling techniques is a prospective avenue that can be explored to enhance the inclusivity and integrity of our dataset.

In the future, We plan to systematically allocate the number of images per age bracket to mitigate biases and ensure that our dataset encapsulates the full spectrum of age demographics. This endeavour provides our commitment to methodological rigour and the pursuit of creating excellent datasets. It upholds strict dataset-creation practices, enabling other researchers to contribute to computer vision and AI.

3.1.2 Annotation of labels of Attire

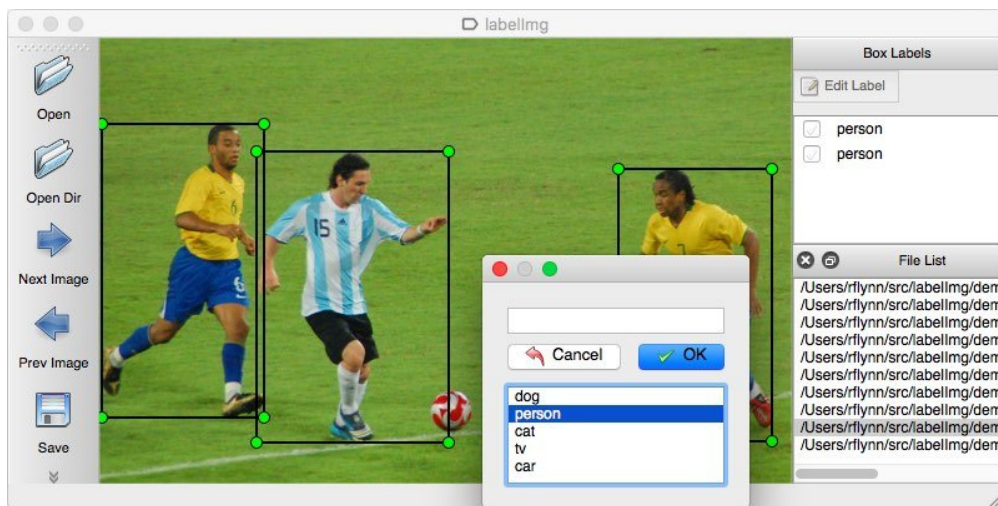


FIGURE 3.3: The figure shows the annotation tool’s simple yet effective interface employed in constructing this dataset.

The first annotation task was to decide on a labelling software (Tzutalin 2015). We only considered free software that gave us annotations in a format that We could use for formats such as PASCAL VOC and YOLO. Regarding the versatility and compatibility of annotations, LabelImg supports various annotation formats, such as machine learning frameworks, by facilitating seamless integration into existing workflows. The software could provide customised shortcuts for annotation and supports image zooming and panning. It also supports collaboration, which ensures that the annotators can effectively collaborate and annotate large datasets (Russell et al. 2008).

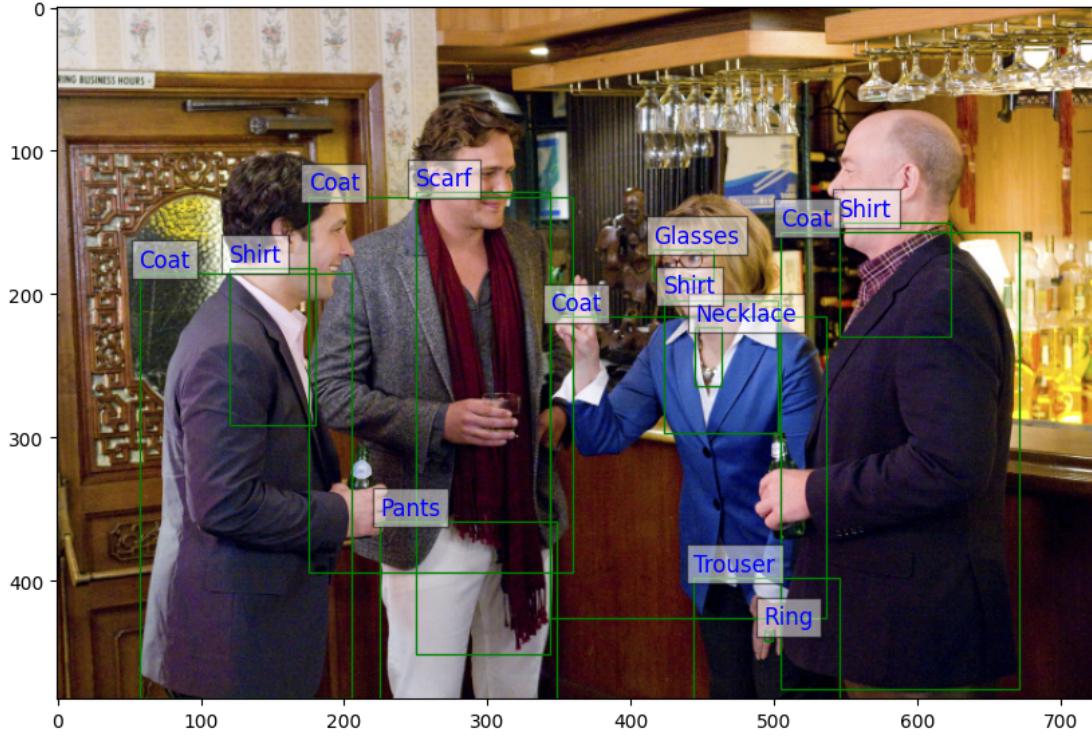


FIGURE 3.4: The image shows a sample of the type of annotation that was performed by the Annotators.

3.1.3 Labelling Attire

Building on our previous work (Yousuf et al. 2019) and responding to the challenges, We assembled approximately 8000 images from the IMDB-wiki dataset. The images were annotated for classification across 60 different attire classes. This resulted in the creation of the IMDB-Attire dataset. Our central objective was to cultivate a comprehensive dataset capable of facilitating entire detection within images that capture from diverse real-world settings, including but not limited to live street surveillance cameras.

The IMDB-attire dataset exhibits a diverse range of image sizes as seen in Figure 3.5, spanning from 22.2 megabytes to 8 kilobytes, reflecting the heterogeneous nature of image dimensions encountered in real-world scenarios. Notably, the smallest image dimension observed within the dataset measures 225×400 pixels, while the largest image attains dimensions of 2547×5320 pixels. This variance in image sizes stresses the importance of robust image processing

techniques to accommodate diverse resolutions and aspect ratios encountered in practical applications.

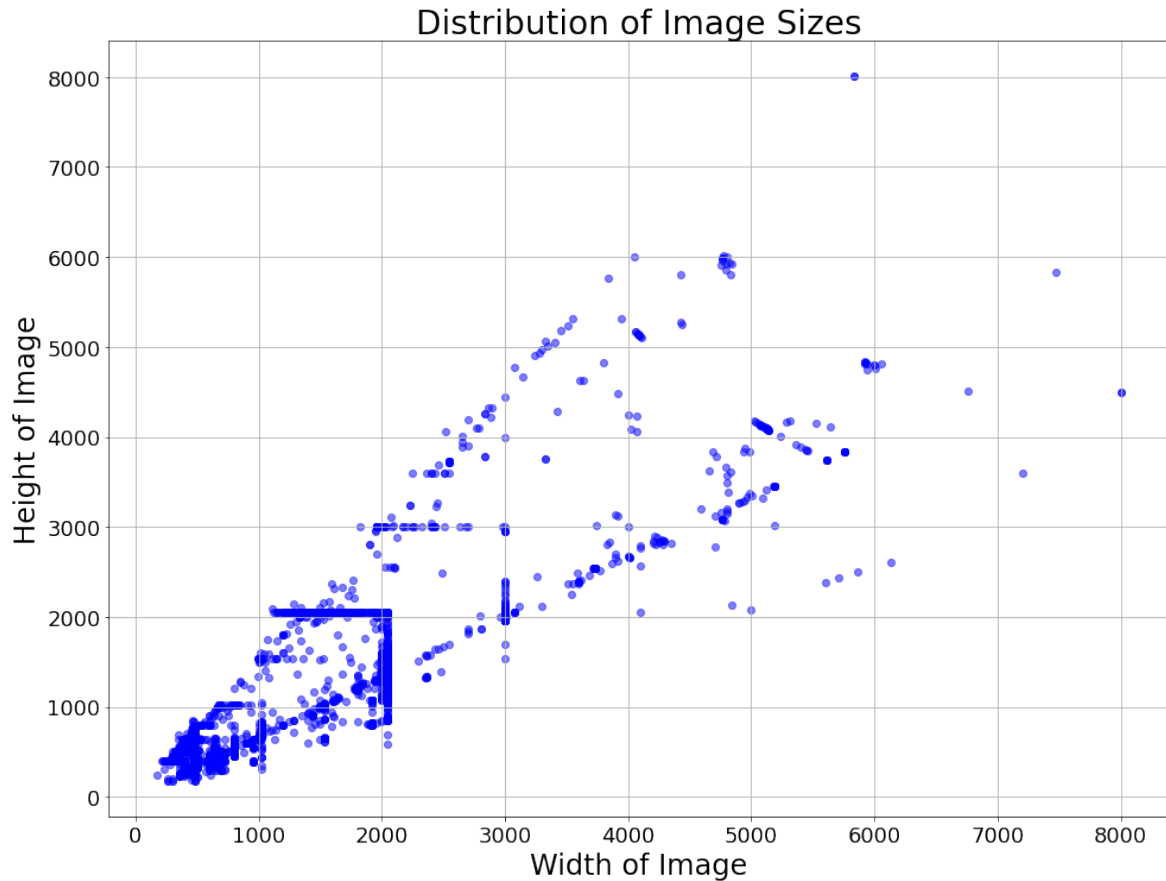


FIGURE 3.5: Graph shows the size and width of every image in the dataset

In addressing the challenge of selecting appropriate labels for annotation within our research framework, We explored various strategies to ensure comprehensive coverage of attire categories. Initially, We established a predefined set of labels, intending to adhere strictly to these classifications. However, We encountered limitations with this approach, as it risked leaving certain attire items unlabelled, particularly those that had yet to be anticipated during the initial label definition phase. In response to this challenge, We adopted a more inclusive approach, wherein We resolved to label any attire item present in the images, regardless of whether it had been previously considered within our predefined set of labels. This encompassed a broad spectrum of items individuals wore, extending beyond traditional clothing articles to

accessories such as watches, bracelets, and culturally significant items like firearms in specific contexts. This iterative labelling process unfolded recursively, progressively adding labels as new attire categories emerged from the image data.

Upon comprehensive annotation, the IMDB-attire dataset revealed insights into the distribution of attire classes across the dataset. Among the 60 labelled classes, "Coat" emerged as the most prominently represented class, with 5,319 instances identified within the dataset. This observation underscores the prevalence and significance of coat-related Attire withinAttireannotated image samples.

The IMDB-attire dataset (Yousuf et al. 2019) represents a concerted effort to curate a diverse and comprehensive collection of images annotated for attire classification. The dataset offers valuable resources for developing and evaluating attire detection algorithms tailored for real-world applications, such as street-level surveillance and live image analysis, through manual labelling and inclusion of images sourced from cinematic contexts. Moreover, the dataset's diverse range of image sizes and attire classes ensures its relevance and applicability across various computer vision research and applications.

3.1.4 Annotations in Detail



FIGURE 3.6: The image shows Earrings in the foreground. Although the background is blurred, a human can see that it is a woman, and she is also wearing Earrings. The Earrings have been labeled.

In this section, We will present some sample images from my dataset. The dataset is an asset to the computer vision community for developing, testing, and priming object detection algorithms for real-world deployment. One of the main highlights of the dataset is that a human can see a face and predict the age and gender. The face has been labelled. All the different attire have been labelled. The dataset's images represent various environments. This includes urban settings, rural landscapes, and indoor and outdoor scenarios, reflecting real-world conditions. Such diversity enhances the dataset's versatility, enabling models to adapt to different contexts and improving their performance across varied situations.

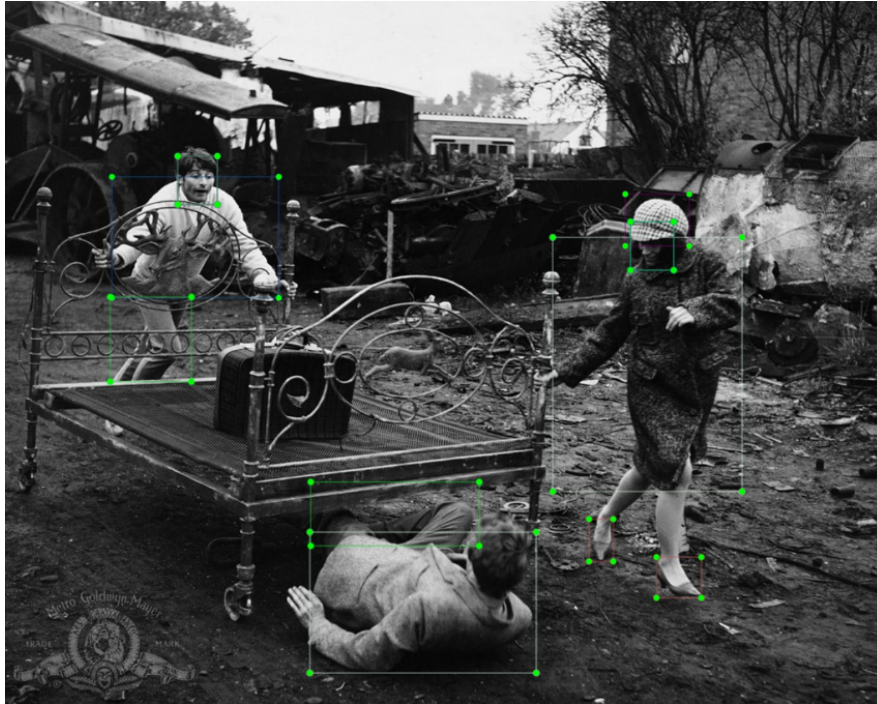


FIGURE 3.7: The image captures a historical, featuring individuals in period-specific clothing. Detailed annotations of different coats (person lying on the ground) emphasise the dataset's capability to handle diverse attire in different settings.

Another aspect of the dataset includes images featuring multiple instances of the same or different objects. This characteristic introduces a layer of complexity that is often encountered in real-world settings. Also, the dataset includes different object sizes, and the same objects are occluded. Objects within the images vary significantly in size, and many are partially occluded, simulating realistic conditions that object detection models must accommodate. This is crucial for training algorithms that are accurate and resilient to the diverse and unpredictable nature of real-world environments.

A selection of images is presented so you can glimpse the dataset. These images visually represent the dataset's richness and diversity, highlighting its potential to support the advancement of object detection technologies. By showcasing the variety of labels, environmental diversity, multiple object instances, and the challenges of object size and occlusion, these sample images highlight the dataset's importance and its contribution to the field of computer vision.

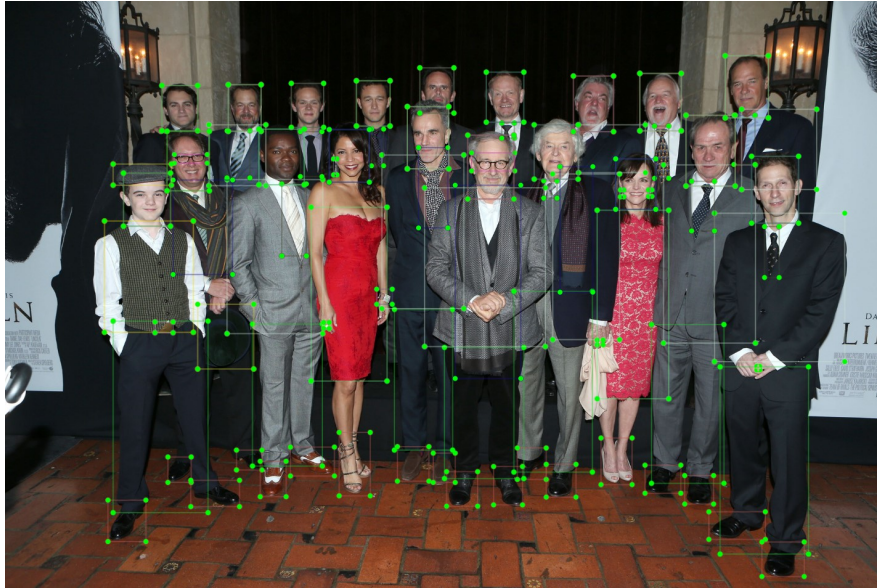


FIGURE 3.8: Depicting a group photo at a formal event. The image highlights the detection of multiple individuals in various attires, ages and genders, demonstrating the dataset’s capability to handle diverse clothing styles and age and gender in group settings.



FIGURE 3.9: Image in a casual indoor setting. The image features three individuals engaged in a conversation, highlighting the dataset’s ability to detect and differentiate human figures in natural, accurate, everyday scenarios. The scene shows the importance of context-aware detection in dynamic environments.

3.1.5 Annotation of Age and Gender

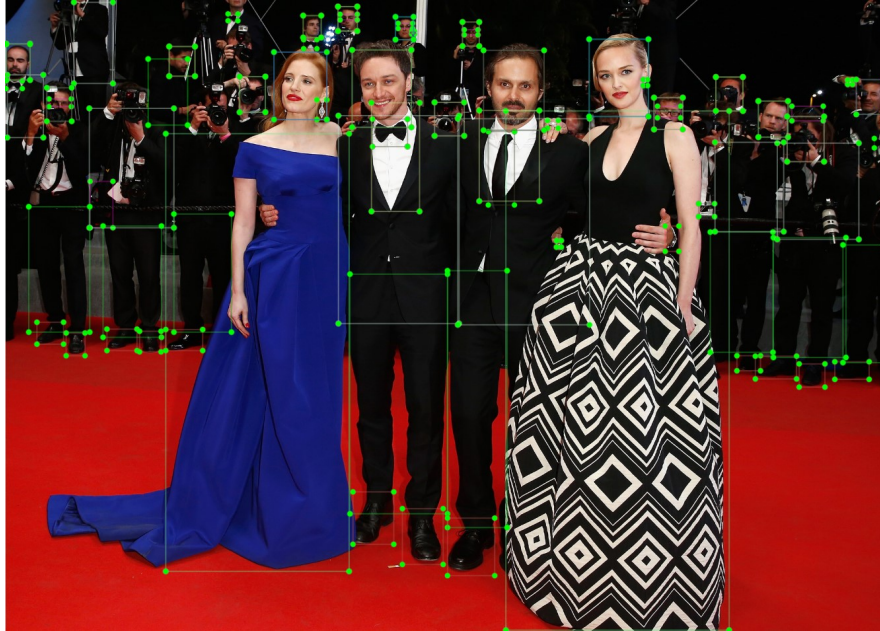


FIGURE 3.10: The faces of two men and two women are clearly visible, all facing forward towards the camera. This image illustrates the dataset’s ability to accurately detect and label facial features, age, and gender in a formal event setting in crowded environments and with high visibility and varied facial expressions.

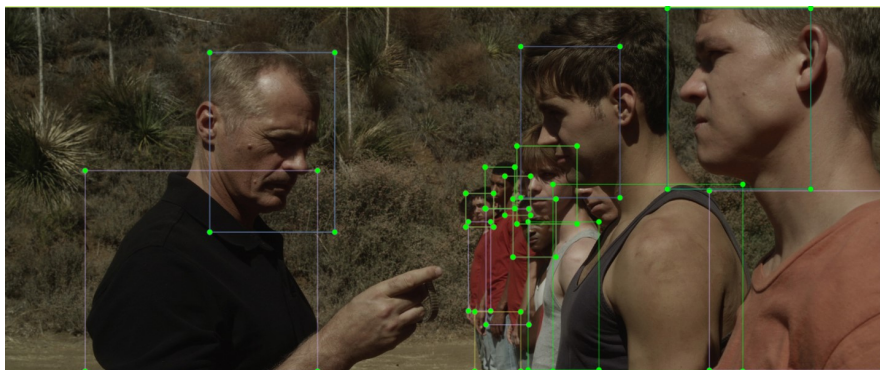


FIGURE 3.11: The image demonstrates the dataset’s capacity to detect and label faces from different angles and in various lighting conditions, emphasising age diversity in a natural environment.



FIGURE 3.12: Image from the dataset showing a diverse group of individuals outdoors. The scene includes men, women, and children, with faces visible from various angles. This image highlights the dataset’s ability to detect and label faces of different ages and genders in a group setting with varied lighting and natural environment conditions.

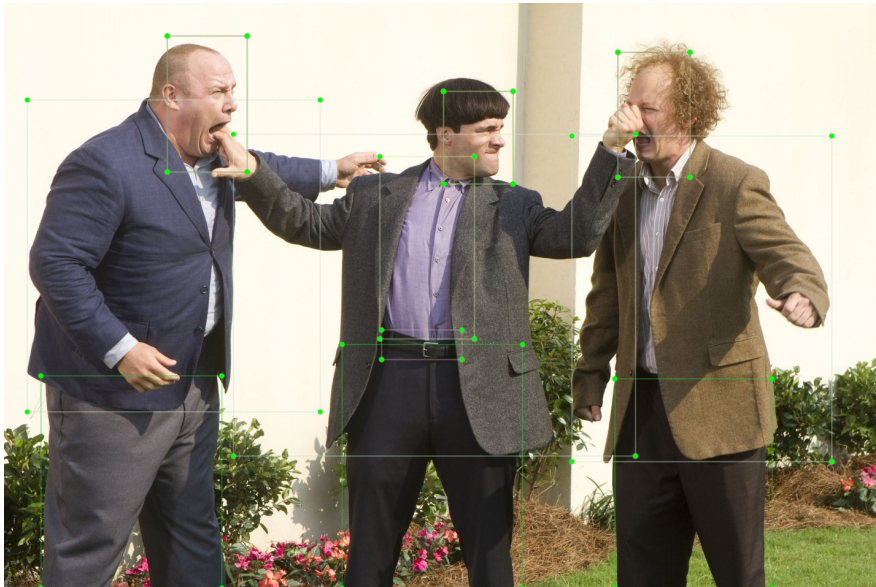


FIGURE 3.13: Image from the dataset featuring three adult men engaged in playful physical natural interaction in an outdoor setting. The men are facing each other, with expressions indicating surprise and amusement. This image demonstrates the dataset’s capability to detect and label faces with diverse expressions and interactions, highlighting the ages and genders of facial features in a dynamic context.



FIGURE 3.14: This image highlights the dataset's ability to detect and label faces of different genders and ages in varied professional contexts, incorporating diverse expressions and attire.



FIGURE 3.15: Image from the dataset featuring a man in a suit sitting in a studio setting, smiling towards the camera. This image demonstrates the dataset’s capability to detect and label faces in professional and media-related contexts. Also, the face on the screen is labelled, showing the dataset’s extensive labelling.

The labelling of age and gender annotations depended on successfully detecting faces and images. We decided to implement and test out Dlib (King 2009), Haar(Soo 2014), OpenCV Dnn modules(Bradski and Kaehler 2000), retina face(Deng et al. 2020), DSFD (dual shot face detection)(Li et al. 2019) and EXTND (extremely tiny face detector)(Yoo et al. 2019) on a small subset of our primary dataset. The objective was to select the best available face detection algorithm. And detect all the faces in our, 8000 images. After comparing these face detection algorithms, We decided to use Retina-net as our initial experiment showed that it was the best algorithm to suit our needs. State-of-the-art DSFD did not outperform retina net in our experiments on our dataset.

After detecting the faces, We performed experiments where We tested existing age and gender estimation algorithms on our test dataset. We tried different size bounding boxes on our face images, including 0% to 60% of the surrounding area around the faces. We saw that most algorithms worked well around the 40% mark, as pointed out by (Rothe et al. 2015).

We created a small test dataset of around 150 images and rigorously tested procedures to assess algorithm performance. We did not have reference ages for the test subjects in a dataset; therefore, our evaluation primarily relied on subjective assessment and the naked eye to visually inspect and judge the apparent ages depicted in the images. This qualitative approach enabled us to gauge the accuracy of the algorithms in accurately estimating ages relative to human perception.

To ensure the consistency and reliability of the annotation process, the annotations were subjected to the validation process involving three different human annotators. This approach enabled us to mitigate potential biases and discrepancies. The final age of the subjects was determined through the consensus of the annotators. This was done after the artificial labelling of age and gender by our chosen algorithm. This collaborative validation process enhances the robustness and accuracy of the annotated dataset. Among the different age estimation algorithms We evaluated, We found that the Retina net trained on the IMDB-wiki dataset exhibited superior performance when applied to our dataset. This also demonstrated the transfer learning in adapting pre-train models to label new datasets. Despite our efforts to maintain accuracy and consistency throughout the annotation process, it is acknowledged that human judgement is inherently subjective, and discrepancies may arise. Comprehensive reviews of the annotated dataset were conducted to ensure accuracy and consistency, with ongoing efforts to refine the annotation process and address any discrepancies that may arise.

CHAPTER 4

Results

First, we present a comprehensive overview of the evaluation metrics utilised in object detection. Metrics such as mean Average Precision (mAP), True Negatives (TN), False Positives (FP), False Negatives (FN), and the F1 score serve as the cornerstone for evaluating detection performance. Precision, recall, and their interplay are explored alongside precision-recall curves. This is crucial for assessing detection models' effectiveness. Towards the end of the chapter, we will discuss the results of our methodology and experiments.

Secondly, we present the results along with the comprehensive analysis. We then discuss our dataset (**RQ1**). Afterwards, we provide detailed statistics on different labels within our dataset, such as Age and Gender distributions, shedding light on the demographic composition of our data. Additionally, we explore the diversity of backgrounds in our images while addressing the challenges encountered during the annotation process across different classes.

We will also discuss some notable findings and insights gained from implementing state-of-the-art algorithms on our dataset. This is followed by an investigation into image enhancement techniques and the effect of different Face to Facial-Peripheral Ratio (**RQ2**) and image enhancement (**RQ3**) on Age estimation and Gender classification of different Age and Gender algorithms. We also offer insights into improving these tasks' accuracy. Subsequently, we delve into Age estimation, contrasting apparent Age with real Age detection.

Finally, we shift our focus to object detection algorithms based on our datasets targeting different classes. We analyse the results of these algorithms regarding Age estimation and Gender classification. We also provide valuable insights into the efficacy across various demographic groups. Finally, we share our training parameters, including mean average

precision; we also discuss and compare the accuracy of different algorithms on our datasets. This also provides valuable insight into how these algorithms estimate Age and Gender in the wild.

4.1 Metrics for Evaluating Object Detection

Different approaches have been developed to address the increasing demand for precise object detection models. When evaluating and updating detection models, we look at two things. First, we look at the coordinates of the bounding boxes, which measures the accuracy of localisation of the bounding box. Secondly, we look at the correct identification of the object that is being detected in that box.

Detection results typically contain different bounding boxes, confidence scores, and associated classes. Object detection in various algorithms can have a different conception when implemented; the standard output format of these files could significantly vary among different detection algorithms. Nevertheless, bounding box representations are usually represented with numbers that identify the box's top left and bottom right coordinates.

Basics of commonly used metrics on which these object detection metrics have been built.

Ground truth: original bounding boxes labelled by annotators on an image with labels of different classes.

True positive(TP): correctly detected Bounding box identified by an algorithm.

False positives(FP): incorrect detection of a bounding box by an algorithm. It means the detector says there is a bounding box at this location. However, there is none.

False negatives(FN): Any ground truth bounding boxes not detected by the object detection algorithm.

True Negative (TN): is not applicable, given the infinite number of bounding boxes not expected in any image.

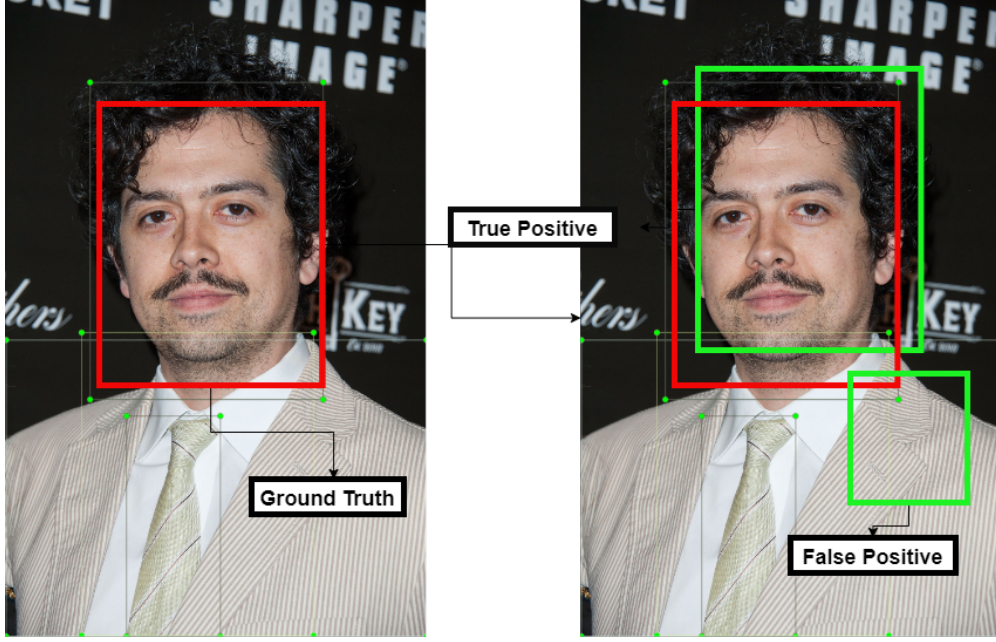


FIGURE 4.1: In the image on the left, you can see the red bounding box as annotated by the annotators. This is the ground truth for the face in this image. You can see the green predicted bounding boxes in the image on the right after the algorithm has been trained and implemented. It is not exactly on top of the ground truth bounding box. You can also see another predicted green bounding box, which shows a false positive.

These definitions rely on the Intersection over Union (IOU), a measure based on the Jaccard Index (Jaccard 1901). IoU calculates the overlapping area between the predicted bounding box ($BoXp$) and the ground-truth bounding box ($BoXgt$) divided by their union. The formula is:

$$J(BoXp, BoXgt) = IOU = \frac{area(BoXp \cap BoXgt)}{area(BoXp \cup BoXgt)} \quad (4.1)$$

A detection is considered correct if $IoU \geq t$ (*threshold*), and incorrect if it is ($IoU < t$).

Due to the absence of True Negatives (TN) in object detection, metrics relying on TN, such as True Positive Rate (TPR), False Positive Rate (FPR), and Receiver Operating Characteristic (ROC) curves, are not commonly employed. Object detection methods are primarily assessed using Precision (P) and Recall (R) (Padilla et al. 2020). These are defined as:

$$P = \frac{TP}{TP + FP} = \frac{TP}{AllDetections} \quad (4.2)$$

$$R = \frac{TP}{TP + FN} = \frac{TP}{AllGroundTruths} \quad (4.3)$$

For object detection, we use Precision and Recall. Precision is the ability of a model to correctly identify objects by representing the percentage of accurate positive predictions. Recall identifies all ground truth bounding boxes indicating the percentage of correct positive predictions from all provided ground truth.

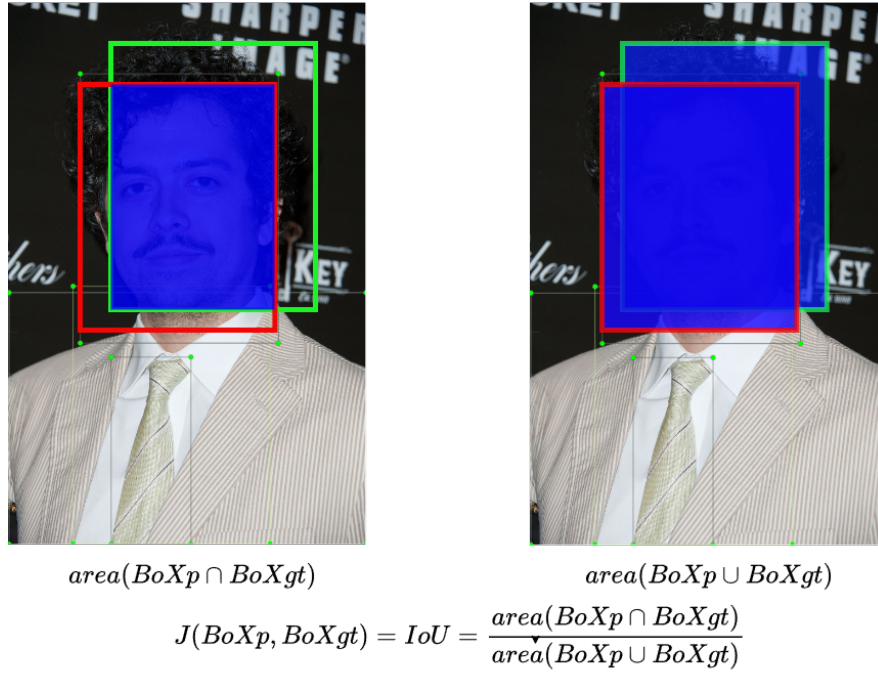


FIGURE 4.2: In the image above, on the left, you can see the intersection of the ground truth bbox and the predicted bbox. We consider the overlap region between the ground truth and the prediction to calculate the intersection area over union. After that, we calculate the area of both the overlapping regions and add to it the area of the non-overlapping region, which is known as the union of ground truth and prediction. This can be seen in the image on the right. After that, we calculate the intersection over union by dividing the areas of both regions.

The Precision-Recall curve illustrates the trade-off between Precision and Recall at different confidence levels associated with bounding boxes generated by a detector. An ideal object detector should achieve high Recall (finding all ground-truth objects) while maintaining high Precision (identifying only relevant objects). A high Area Under the Curve (AUC) on the Precision-Recall plot suggests high Precision and Recall.

However, the Precision-Recall plot often exhibits a zigzag pattern in practical scenarios, making AUC measurement challenging. Two common methods address this limitation: (i) the 11-point interpolation and (ii) all-point interpolation. These two methods help smooth the curve and enhance the accuracy of AUC estimation (Padilla et al. 2020).

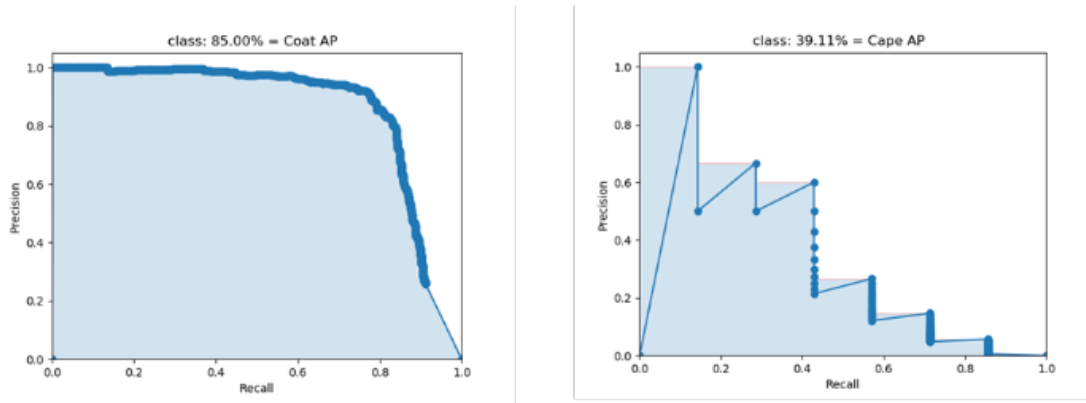


FIGURE 4.3: This is one of the results of our training. The left graph shows the Precision and Recall of the label Coat, and the graph on the right shows the label Cape. The graph on the left shows a high Recall and Precision with AP of 85%. On the other hand, the one on the right shows the zigzag pattern discussed above and AP of 39.11%.

The relationship between Precision and Recall in object detection gives rise to two more concepts of quantifying the quality of object detection. The first is Average Precision (AP), and the second is mean Average Precision (mAP).

Average Precision for 11-point interpolation:

$$AP_{11} = \frac{1}{11} \sum_{R \in \{0, 0.1, \dots, 0.9, 1\}} P_{interp}(R) \quad (4.4)$$

$$P_{interp}(R) = \max_{\tilde{R}=\tilde{R}_{\geq R}} P(\tilde{R}) \quad (4.5)$$

All point Average Precision:

$$AP_{all} = \frac{1}{11} \sum_n (R_{n+1} - R_n) P_{interp}(R_{n+1}) \quad (4.6)$$

$$P_{interp}(R_{n+1}) = \max_{\tilde{R}=\tilde{R}_{\geq R}} P(\tilde{R}) \quad (4.7)$$

The mean Average Precision (mAP) serves as a metric to evaluate the overall accuracy of object detectors across all classes in a particular database. It is calculated by taking each class's average of the Average Precision (AP) values.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (4.8)$$

Here, AP_i represents the Average Precision for the i -th class, and N is the total number of classes being evaluated. The mAP is the average of these individual AP values.

4.2 Dataset construction

Developing a dataset to cater to real-world environments posed significant challenges, as the real world is very dynamic and diverse and comprises spanning wilderness, urban streets, cluttered areas, and poorly lit places. Despite recent advances, the practical applications of these algorithms require extensive manual image annotations (Bradski 2000). Attire mirrors cultural norms and social standings and conveys personal attitudes and brand affiliations. This attire recognition is useful in surveillance, fashion, and retail analytics.



A: Original Image without any image Enhancement



B: EDSR based Image Enhancement



C: GFP-GAN based Image Enhancement

FIGURE 4.4: A: Original Image. B: Effects of EDSR on Image (Lim et al. 2017). C: GFP-GAN (Wang et al. 2021)

4.2.1 Image Enhancement Techniques

Single-image super-resolution (SISR) has garnered substantial attention due to its ability to enhance image details and quality. We focus on two particular works, “Enhanced Deep Residual Networks for Single Image Super-Resolution” (EDSR) (Lim et al. 2017) and “Towards Real-World Blind Face Restoration with Generative Facial Prior” (GFP-GAN) (Wang et al. 2021).

Research (Wang et al. 2018) shows that image enhancement techniques can play an important role in improving the quality of low-resolution and blurry images. Enhancing our images can improve the accuracy of Age estimation and Gender classification, as they heavily rely on facial features such as wrinkles in texture and facial contours. These features are not that prominent in low-resolution images, which could lead to low performance in Age estimation and Gender classification. Image enhancement, particularly techniques like SISR (e.g., EDSR), sharpens these features, making them more distinguishable and allowing the model to make more accurate Age and Gender predictions. It also helps in the restoration of facial images that might have been degraded due to noise, blurriness or other distortions. Using super-resolution techniques can restore important facial features, resulting in better detection results.

Image enhancement removes biases that may arise from varying image quality across different age demographic groups. For example, if a dataset contains lower-quality images from a certain age group, the model might be biased against accurately predicting that group’s age. Enhancing all images to a consistent quality level ensures the model generalises better across all age groups, leading to more accurate age predictions (Huang et al. 2017).

The application of the two techniques (GFP-GAN & EDSR) is quite different from each other. We aimed to examine how the various algorithms will enhance Age estimation accuracy. It is essential to remember that the EDSR is focused on improving the quality of low-resolution images. In contrast, the GFP-GAN is more focused on restoring blind faces, such as images that have gone under degradation or noise due to storage or elements. GFP-GAN is primarily designed to restore facial features in such images. EDSR is focused on learning how to

map low-resolution features to high-resolution targets. GFP-GAN has been trained on many facial images to encapsulate prior knowledge of the intrinsic characteristics of human faces, enabling the model to restore facial details and features without prior knowledge of the degradation process. It also incorporates techniques like adversarial training and perceptual loss to improve the quality of the image restored. Compared to EDSR, GFP-GAN was better suited to our needs, and a few initial tests proved that, as seen in Figure 4.4. GFP-GAN provided a better definition and information on facial images, as it was already trained to address facial image enhancement.

One of the objectives was to enhance images **RQ3** so that age estimation algorithms could estimate age as closely as possible to that of human counterparts. We focused on state-of-the-art algorithms to achieve this goal. We narrowed down the two best algorithms for the job and then tested many Age estimation algorithms with many iterations of different scales of the face image. After visually assessing the results of the best setting, we finalised the image enhancement process. The significant effects of image enhancement on Gender identification have been due to the biases in the IMDB dataset. Our experiments showed that GFP-GAN had a far more significant impact on improving the accuracy of both Age and Gender, as shown in Figure 4.4.

4.2.2 Dataset Statistics

We chose the IMDB-WIKI Dataset, which stands out as a renowned resource for Age estimation and Gender-based research. Created by scholars at ETH Zurich, this dataset comprises images sourced from IMDB.com, spanning movie scenes, backstage glimpses, and real-life moments of over 100,000 actors across their lifetimes. In total, the dataset encompasses 460,723 comprehensive images. A pertinent query arises: why choose this specific dataset? Would it not be more efficient to define labels and then scour the internet accordingly for targeted images? The rationale behind our choice lies in our research's broader implications and practical applications. We chose this dataset because of the availability of different Age and Gender labels. After labelling Attire, our job of labelling ages and genders would have been far easier than if we scrapped images from any other source.

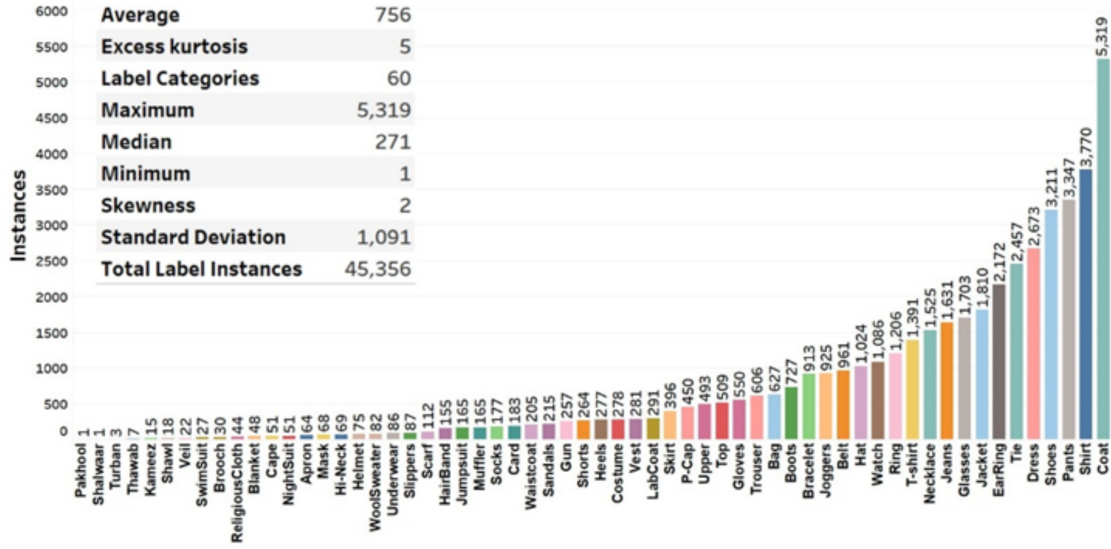


FIGURE 4.5: Bar graph showing statistics of 60 labelled classes in our IMDB-Attire dataset (Yousuf et al. 2019).

We initially collected 3,500 images for our preliminary experiments, which later expanded to 8000. Termed the IMDB-attire dataset (Yousuf et al. 2019), our selection was driven by its diverse array of scenarios. The movie’s multifaceted settings provided a rich tapestry for our research pursuits. We explored various approaches to address the challenge of selecting labels for annotation. Initially, we opted to predefined and adhere to a set of labels. However, this approach posed limitations, as certain things not previously considered would remain unlabelled. Consequently, we pivoted to labelling anything classified as attire, encompassing items worn by individuals beyond conventional clothing, such as watches, bracelets, and culturally significant accessories like guns in specific contexts. Employing a recursive labelling process, we continually augmented the label set as new attire categories emerged in the images. After labelling 8000 images, our dataset comprised 60 distinct classes.

4.2.3 Face Detection Results

We considered both contemporary state-of-the-art methods and classical deep learning approaches to choose the best face detection algorithm. After exploring the latest face detection algorithms, we chose DSFD (dual shot face detector) (Li et al. 2019) and EXTD (Yoo et al. 2019)(Extended tiny face detector), as both of these algorithms were ranked very highly on face detection accuracy. We also looked at deep learning algorithms such as Retina net (Zhang et al. 2020) and MTCNN (multitasked convolution neural network) (Xiang and Zhu 2017). For comparison, we also looked at SSD (single-shot multi-box detector). We also investigated four classical OpenCV-based detectors (Bradski 2000) to assess their efficacy on our dataset. We curated a subset consisting of carefully selected 30 images from our main dataset, manually annotating each image to establish a ground truth of 136 faces.

Name of Algorithm	Faces detected	False Positive	False Negative
DSFD (Li et al. 2019)	159	14	9
EXTD (Yoo et al. 2019)	141	2	9
MTCNN (Xiang and Zhu 2017)	124	1	27
RetinaNet(Zhang et al. 2020)	132	0	4
DLIB (King 2009)	74	0	57
OpenCV(HAAR) (Soo 2014)	62	3	67
SSD(Liu et al. 2016c)	51	0	73
Ground Truth	136		

TABLE 4.1: Comparison of different face detection Algorithms on our dataset

The outcomes of our experiments are provided in the table above. After our experiments and quiet deliberation, we identified the Retina-Net as a primary face detection algorithm. Its superior performance detected 132 faces out of the 136, with only four false positives. DSFD and EXTD also performed well; however, many false positives overshadowed their performance. The state of the art of algorithms also failed to capture many Faces. Retina-Net, only missed around four faces, which was extremely hard to detect. On top of that, the Retina-Net did not have any false positives.



FIGURE 4.6: Image A and B are results from DSFD (Li et al. 2019), algorithm performance varies as it fails to detect side poses of the face; it effectively detects an occluded face. ExtD (Yoo et al. 2019) and Retina-Net (Zhang et al. 2020) miss the same face as shown in Figure F and D. Haar (Soo 2014) are also not acceptable, as shown in Figure C.

As we can see in Figure 4.6, we can see the detection capabilities of DSFD, EXTD, HAAR, and Retina-Net. A and B are DSFD results, and C and D are EXTD Results. Our results show that Retina net performs better than the state-of-the-art algorithm DSFD, as can be seen by comparing images A and F. Further experiments are required to assess the claims made in the DSFD paper. As this is not the scope of our research, we will leave this for later.

4.2.4 Face Alignment Results

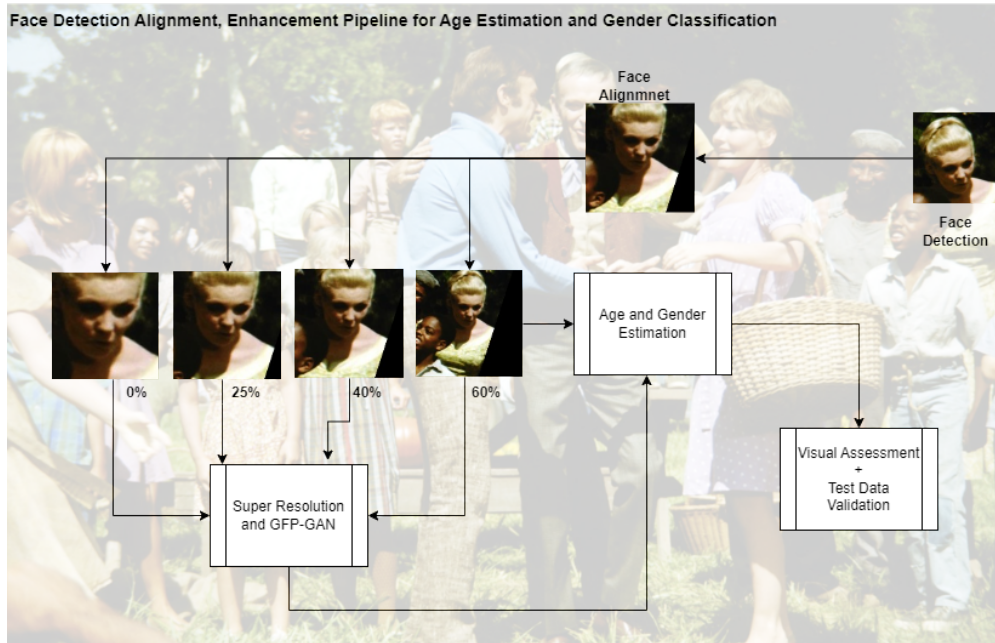


FIGURE 4.7: The image shows how the face is detected in images and then aligned. Afterward, a different ratio of the area around the face is selected, and image enhancement techniques are applied. These images are then tested on various Age Estimation and Gender Classification Algorithms.

The main purpose of face alignment is to ensure that the input faces in the images are of consistent size, centered, and aligned to maintain a central facial position of all faces with minimal background interference. There are many approaches to do this, from intricate alignment techniques involving precise facial landmark detection and image wrapping to applying face transformations to make sure that the nose of each face is in the same location and the eyes are also in the same location, and so on.

While constructing our dataset, we faced the challenge of working with different datasets for Age estimation and Gender classification. We also came across different algorithms trained on these datasets. Most of the faces in these images were in the natural environment, and in most datasets, people's faces were in uncontrolled positions rather than centered frontal poses. Through extensive experimentation, it was evident that each algorithm operated by its pipeline, unique to its dataset and tailored to the dataset it handled. If we wanted to

reproduce the same results, we had to follow the pipeline exactly. Initially, we decided to develop and integrate our facial alignment technique into our solution. The Age estimation and Gender classification algorithms were applied with disappointing results after applying our facial alignment technique to the dataset. Recognising our shortcomings, we opted for the pre-processing pipelines established by individual researchers for their respective algorithms. This decision showed some improvement in the accuracy of the results of different estimation algorithms on our dataset. Finally, we decided to use the same algorithm and pre-processing pipelines because the dataset was a subset of the IMDB-Wiki dataset. Naturally, it produces the best results when predicting Gender and Age labels on the IMDB attire dataset. Building upon this insight, we integrated the robust Age alignment technique into the processing pipeline.

Visual assessment was a crucial validation step in our methodology, affirming the efficacy of integrating and establishing Age alignment techniques in our framework. These empirical observations highlighted the importance of visual value in evaluating the performance of estimation algorithms. The robust face alignment procedure diverged from conventional methods focusing on pixel-perfect precision. Our experiments demonstrated that it is superior to the lay and Mark-based approaches. This is also apparent from the LAP challenge (Ponce-López et al. 2016), where the failure rate is around 1% compared to landmark techniques' 5% failure rate.

4.2.5 Training Charts

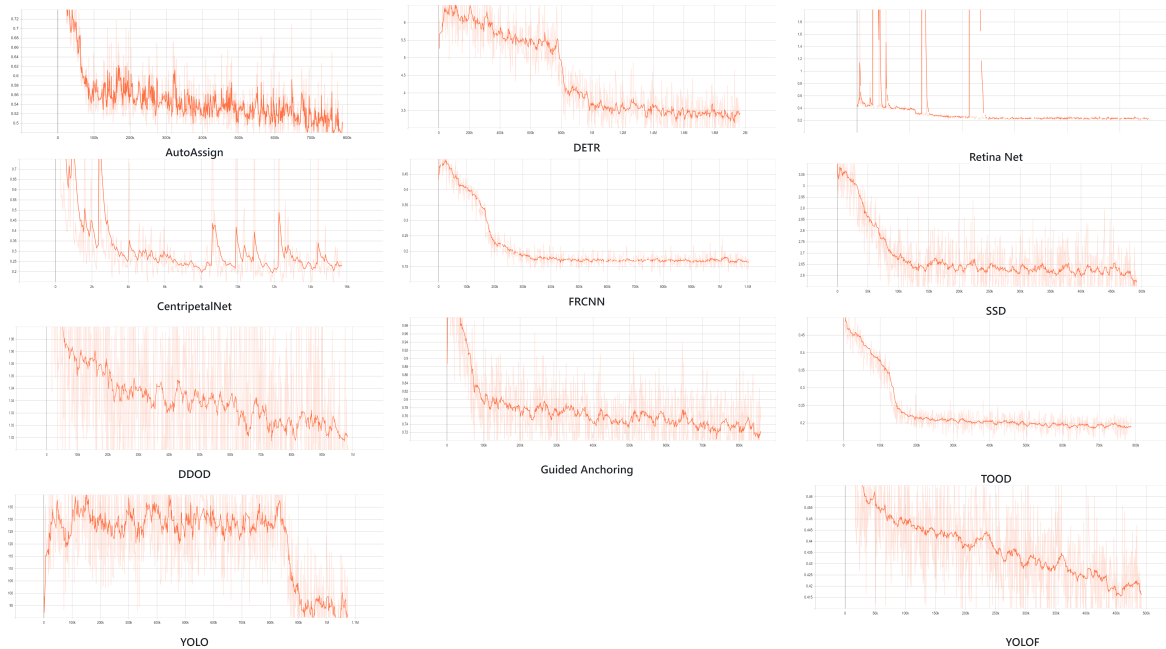


FIGURE 4.8: Training loss curves for various object detection models. The graphs depict the loss convergence over iterations for AutoAssign, DETR, RetinaNet, CentripetalNet, FRCNN, SSD, DDOD, Guided Anchoring, TOOD, YOLO, and YOLOF. Each plot shows the model’s performance during training, highlighting differences in learning rates and stability across different architectures.

After implementing state-of-the-art algorithms on our datasets and looking at the specific components of contemporary object detection algorithms, we critically examined the respective strengths and limitations. We looked at the loss function, which consists of two types in object detection: the classification loss, typically represented by cross-entropy loss, and the regression loss. Cross-entropy loss, which is commonly employed, tends to overlook the inherent class imbalance in most datasets, prompting researchers to investigate alternative loss functions, such as focal loss, to address this issue. Progressing the bounding box coordinates helps minimise the disparity between the ground and predicted bounding boxes.

We also looked at prediction methodologies, which looked at two predominant approaches: Anchor box prediction and key point-based bounding box prediction. The anchor-based prediction used by algorithms such as SSD (Liu et al. 2016c) and YOLO (Jiang et al. 2022) uses presets and boxes where we expect the object to be present in the images, which prolongs

the training time; however, they are quicker regarding prediction time. Another issue is that anchor sizes must be customised according to the dataset. In comparison, key point prediction, which is used by algorithms such as *deter*, which are transformer-based, tries to address the limitations associated with anchor-based strategies. Transformer-based detectors, notably exemplified by DETR (Carion et al. 2020b) and analogous approaches, represent a paradigm shift in object detection, leveraging key point-based methodologies to achieve breakthroughs in accuracy and efficiency. DETR, along with its counterparts such as Center-Net (Duan et al. 2019), embodies the transformative potential of this approach.

4.2.6 Results of Face to Facial-Peripheral Ratio

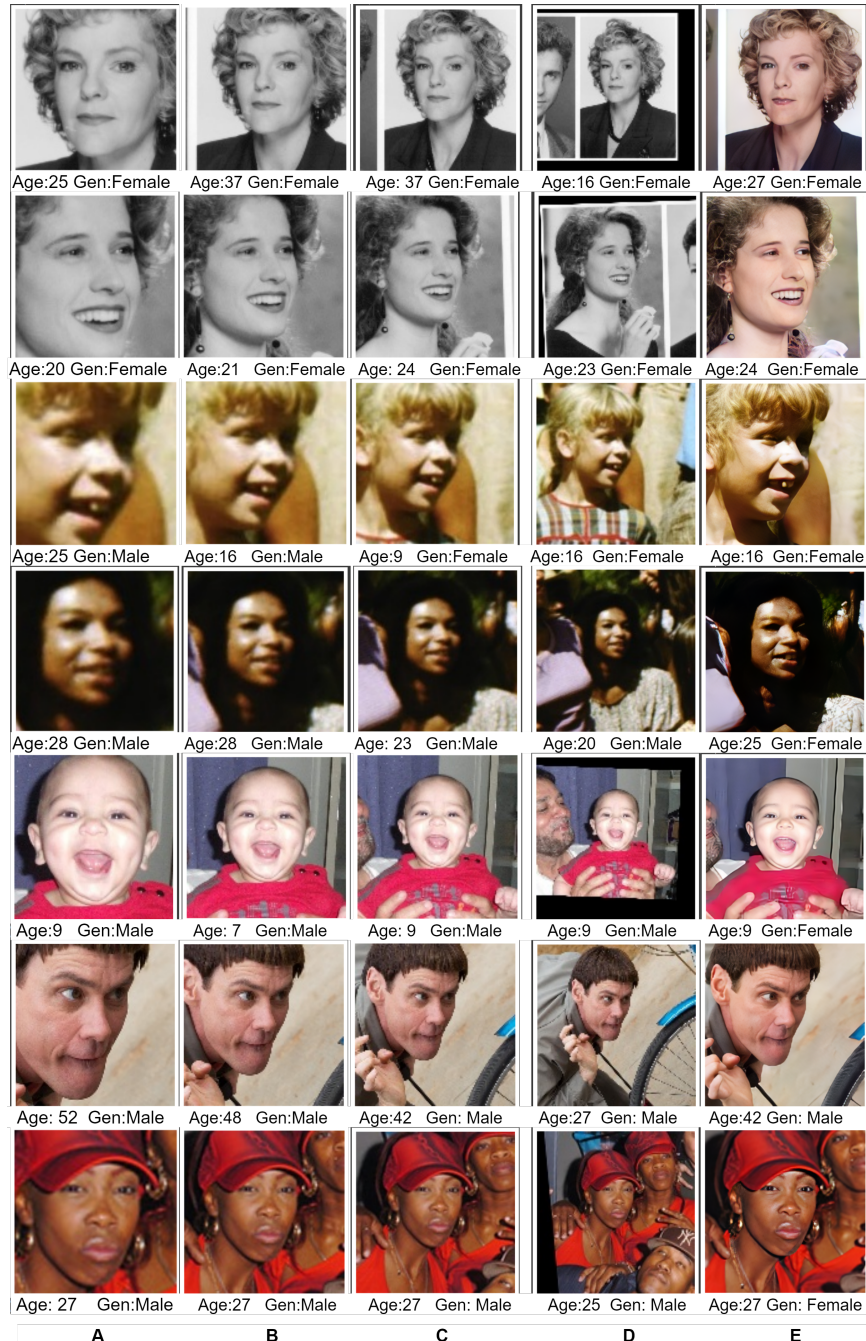


FIGURE 4.9: The image compares different ages at different area ratios around the face. We can see a clear improvement in Age and Gender results after selecting 40% of the area around the face. Column C and a further improvement can be seen in column E, where super-resolution is applied to the face image.

In this section, we discuss **RQ1** and **RQ2** by referring to Figure 4.9, containing sample figures generated during the creation process of our dataset. We address the impact of incorporating different percentages of area around a face and applying image enhancement on images to see their effects on the results of deep learning algorithms trained to predict age and gender. Figure 4.9 contains faces of varying ages, incorporating different percentages of the area around the face, before inputting it into the age estimation algorithm. Figure 4.9 compares results of age and gender on facial images with different face area ratios with and without image enhancement techniques.

Our second research question focuses on how incorporating different percentages of areas around the face helps deep learning algorithms predict age and gender. The sample images we provided show that incorporating the area around the face before passing the image to age estimation and gender classification algorithms can greatly affect prediction accuracy, as shown in Table 4.2.

In Figure 4.9, we can see that in column A, around 0% of the area around the face has been incorporated. Column B incorporates around 25% of the area around the face. Column C incorporates around 40% of the area around the face. Column B incorporates 60% of the area around the face in the wild. Column E is the same as column C but with an image enhancement technique, the GFP-GAN algorithm.

Age prediction is arguably acceptable in columns A and B, where the minimal area around the face has been incorporated. However, the Gender classification is very poor, as many females have been classified as males. The context around the facial features is limited, and the algorithm struggles with ambiguous features.

In column D, we see that around 40% of the area around the face has been incorporated, and we noticed that a remarkable improvement in both Age and Gender prediction is observed. We can conclude that incorporating more contextual information around the face helps these deep-learning algorithms predict age and gender more accurately. The algorithm looks at relevant features such as clothing earrings and other background elements, contributing to a more accurate prediction of Age and Gender.

When we look at column D, where around 60% of the area around the face has been incorporated, we still get predictions, but with diminishing returns. There is a balance around how much area we should incorporate around a face to get the best results in the wild. The study suggests that too much area around the face might introduce more noise than useful context, and the algorithm's ability to estimate age and gender accurately is weakened. The dataset demonstrates that around 40% of the surrounding area around a face provides a balance between contextual information and relevant facial features.

4.2.7 Image Enhancement Results on Age and Gender Prediction

Our third research question (**RQ3**) was how image enhancement techniques such as super-resolution and GFP-GAN affect apparent Age and Gender estimation. Our experiments show that image enhancement techniques, especially Generative Facial Prior GAN, are important in improving the apparent age and gender estimation of faces. Suppose you look at column C and column E in Figure 4.9. In column C, we have around 40% of the area around the face, and column E has the same face again where GFP-GAN has been applied. We can see that in column E images, because of the application of GFP-GAN, there is a significant improvement in the clarity and detail of facial images. Some black and white images have also been coloured because of GFP-GAN. Especially when we look at different races, such as images of females of African descent, who are usually classified as males. However, applying image enhancement techniques helps deep learning algorithms correctly classify them as females. In our experiments, we also saw that image enhancement techniques helped Age estimation and Gender classification algorithms classify faces in the background and also helped identify images of faces that were blurry or unfocused. It is reasonable to infer that such advanced enhancement techniques further boost prediction accuracy. GFP-GAN can generate realistic, high-resolution facial details, enhancing the model's ability to interpret age-related features accurately and reduce errors in gender classification.

Concluding, approximately 40% of the surrounding area around the face provides a decent balance of contextual information and facial features. At the same time, super-resolution

techniques such as GFP-GAN can also affect the accuracy of the prediction of deep learning-based age and gender estimation algorithms.

4.2.8 Age and Gender Results on state-of-the-art algorithms

Method	Top 5 Ages	Gts	Dets	Recall	AP	mAP_5
Guided Anchoring	52	102	821	0.951	0.344	0.310
	42	69	614	0.971	0.312	
	25	167	1034	0.898	0.305	
	40	47	849	0.787	0.294	
	24	51	534	0.922	0.294	
Auto Assign	63	8	50	0.5	0.837	0.782
	40	69	460	0.87	0.829	
	52	102	736	0.804	0.761	
	20	123	823	0.813	0.753	
	24	51	474	0.843	0.733	
DETR-Deform	7	2	10	0.5	0.5	0.415
	40	7	216	1	0.443	
	57	19	239	0.947	0.407	
	42	69	589	0.957	0.386	
	52	103	838	0.961	0.339	
DETR	57	19	53	0.737	0.53	0.406
	13	2	4	0.5	0.5	
	40	8	15	0.5	0.361	
	40	7	50	0.571	0.35	
	17	2	28	1	0.292	
DDOD	42	69	830	0.899	0.234	0.221
	25	167	1084	0.784	0.227	
	40	7	268	1.000	0.330	
	20	123	997	0.764	0.164	
	52	102	1103	0.833	0.151	
FRCNN	57	19	39	0.789	0.498	0.377
	24	51	160	0.725	0.361	
	42	69	131	0.638	0.36	
	52	103	220	0.66	0.334	
	40	167	416	0.653	0.332	
SSD	42	69	498	0.928	0.322	0.289
	57	19	117	0.895	0.291	
	24	51	426	0.922	0.29	
	28	76	595	0.882	0.275	
	40	7	94	0.714	0.269	
TOOD	63	8	91	0.75	0.469	0.402
	57	19	109	0.895	0.412	
	40	69	280	0.899	0.406	
	52	102	555	0.971	0.382	
	37	96	608	0.885	0.342	
YOLOv3	40	6	16	0.333	0.333	0.306
	52	103	430	0.786	0.325	
	57	19	83	0.737	0.303	
	20	124	531	0.79	0.289	
	42	69	231	0.71	0.28	
YOLOF	63	8	131	0.625	0.357	0.313
	40	102	811	0.892	0.322	
	42	69	553	0.928	0.314	
	20	123	1004	0.854	0.3	
	24	51	603	0.922	0.271	

TABLE 4.2: mAP of top 5 age labels compared against different methods.

This section focuses on answers (**RQ1**) and shows the results of state-of-the-art algorithms trained on our datasets. We found that most of the algorithms could cater to the diverse nature of the problem. However, some algorithms, such as DDOD, failed to accommodate all the different labels as shown by table 4.2. Similar to attire labels, we evaluated the efficacy of the dataset on various object detection algorithms to see how these detectors would perform under such a complex task. We also wanted to see the dataset’s effectiveness in the real world. For Age, we assessed each of the algorithms based on recall, average precision, and mean of every position in the top five ages, enabling us to comprehensively compare the performance of these object detectors on our dataset.

The object detected that stood out the most was Auto assigned, which demonstrated a performance of 0.782 on the top 5 edges predicted on the validation dataset. Remember that the detector was simultaneously trained on both agenda and 60 different attire labels. Auto-Assign showcased significant potential for real-world applications.

Transformer base algorithms such as DETR-Deform and DETR also emerged as competitive performances by consecutively achieving a mean precision of 0.415 and 0.406. DDOD (Chen et al. 2021), FRCNN (Liu et al. 2016d), SSD, TOOD, YOLOv3, and YOLOF. These methods present varying performance levels in Age estimation, with mAP scores ranging from 0.221 to 0.402. Notably, TOOD (Feng et al. 2021a) stands out with remarkable recall and average precision values, suggesting its potential suitability for Age estimation tasks requiring high precision and recall.

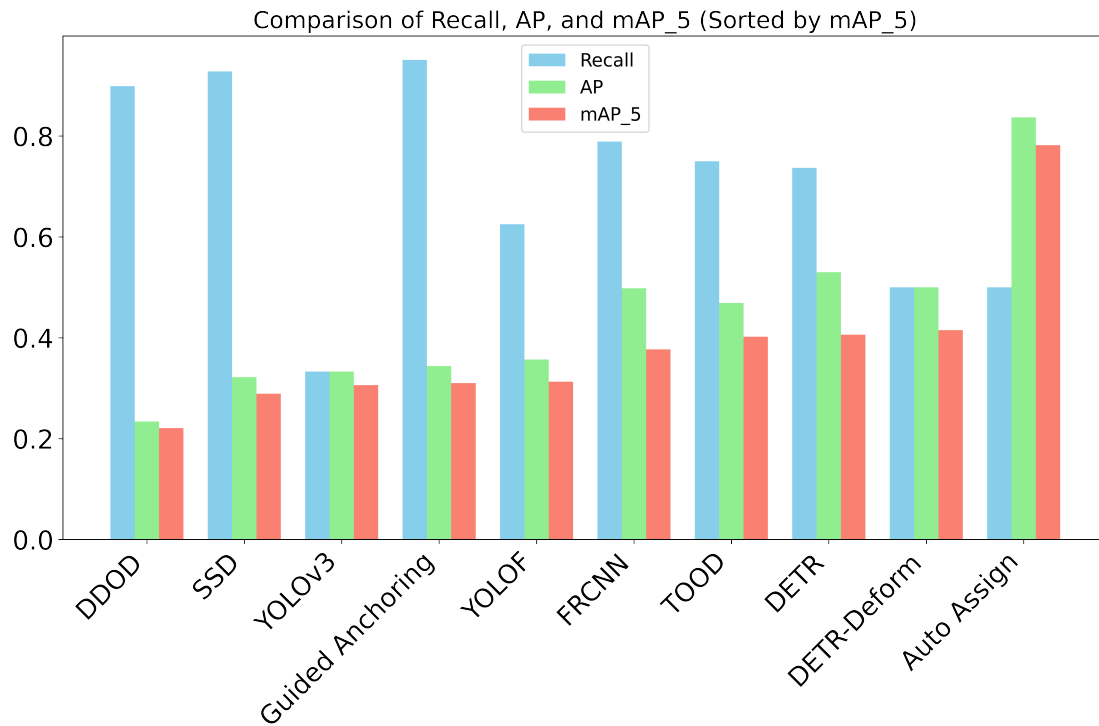


FIGURE 4.10: The Chart gives a visual comparison between the different Algorithms. It compares Recall, Average Precision and Mean Average precision of the top 5 Ages in each Algorithm.

The chart shows a clear ranking among the algorithms. We can see that in the case of Age estimation, Auto Assign (Zhu et al. 2020a) outperforms all the other algorithms when it comes to mAP and AP. DDOD algorithm performed the worst in terms of Average precision, however it had one of the highest Recall. This indicates a gap between its retrieval capability and precision.

One thing that stood out was how different algorithms achieved recall and average precision on different ages on the same dataset. Almost all the labels in the top five categories were identical. These findings lay the groundwork for future research endeavours to advance Age estimation techniques and improve their applicability in real-world scenarios.

4.2.8.1 Mean Absolute Error of Ages on facial Images

Method	MAE	% faces detected
AutoAssign	6.536	66.667
SSD	6.63	62.879
FRCNN	6.852	98.485
DETR	6.901	97.727
DETR-Deform	6.971	82.576
Guided Anchoring	7.309	76.515
YOLOv3	7.325	61.364
YOLOF	7.593	84.848
Retina Net	7.678	90.909
TOOD	7.876	97.727
DDOD	11.709	85.606

TABLE 4.3: Mean absolute error (MAE) between the human-annotated apparent ages and the predicted ages from different methods on the test set. The percentage of faces detected is also on the test set.

The table 4.2.8.1 offers a deeper insight into these algorithms' performance. The analysis also provides validity to the dataset, and different algorithms are trained on it. The MAE values across different methods range from 6.536 to 11.709, reflecting the degree of deviation between the predicted ages and the human-annotated apparent ages. AutoAssign, SSD, FRCNN, DETR, and DETR-Deform demonstrate relatively lower MAE values, indicating higher Age estimation accuracy than other techniques.

Overall, Auto-assign stands out with the lowest mean absolute error of 6.536; however, if you look at the face detection percentage on the test data, it has one of the lowest, 66.667%. This shows us that although it has a low mean absolute error, the algorithm itself doesn't perform well on the dataset as the number of faces it detects is very low. The highest mean absolute error was presented by DDOD of 11.709. FRCNN of 98.485% achieved the highest accuracy when it comes to face detection. It also achieved a mean absolute error of 6.852. So, regarding Age estimation in the wild, Faster RCNN outperformed even transformer-based algorithms on our dataset.

4.2.8.2 Gender Precision on Dataset

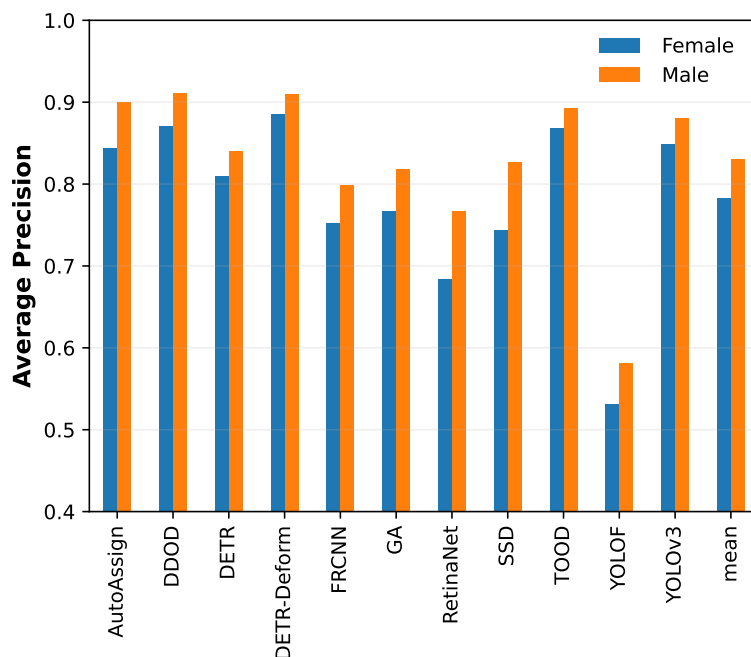


FIGURE 4.11: Average precision of different algorithms to predict the gender on our dataset.

Figure 4.11 compares the average precision of different algorithms in predicting the gender on our dataset. Notably low average precision is exhibited by YOLOF for both male (0.58) and female (0.53) genders. This could be attributed to various factors such as the algorithm's model complexity, feature requirements, and general prediction tuning. Data augmentation, feature engineering, and other hyper-parameter fine-tuning can address this slow average precision on our dataset for YOLOF. When detecting and predicting female genders, DETR-Deform outperforms every other algorithm. DETR-Deform exhibits an average precision of 0.89 for female and 0.91 for male. Compared to DETR-Deform, DDOD has the same average precision in predicting the male gender, but a slightly lower average precision (0.87) when predicting the female gender. DETR and guided anchoring methodology are closest to the mean of combined prediction capabilities of all the different algorithms.

4.2.8.3 Attire Detection Results on the state-of-the-art algorithms

Method	Top 5 Classes	Gts	Dets	Recall	AP	mAP_5
Guided Anchoring	Coat	348	1181	0.948	0.854	0.785
	Tie	161	528	0.882	0.828	
	Hat	54	245	0.907	0.777	
	Jeans	111	475	0.919	0.753	
	Dress	178	895	0.916	0.713	
Auto Assign	Tie	161	872	0.919	0.837	0.765
	Coat	348	1952	0.957	0.829	
	Jeans	111	656	0.883	0.761	
	Glasses	110	688	0.845	0.753	
	Hat	54	351	0.889	0.733	
DETR-Deform	Tie	161	508	0.925	0.871	0.8336
	Coat	348	1340	0.931	0.854	
	Jeans	111	594	0.883	0.779	
	Glasses	110	525	0.855	0.779	
	Dress	178	1218	0.904	0.748	
DETR	Coat	348	5052	0.934	0.835	0.805
	Tie	161	541	0.839	0.794	
	Glasses	110	414	0.791	0.747	
	Jeans	111	774	0.829	0.71	
	Dress	178	2855	0.882	0.706	
DDOD	Coat	348	1275	0.931	0.819	0.785
	Tie	161	595	0.888	0.8	
	Jeans	111	552	0.91	0.759	
	Dress	178	967	0.91	0.668	
	Hat	54	304	0.87	0.661	
FRCNN	Coat	348	482	0.899	0.843	0.7414
	Tie	161	181	0.857	0.814	
	Glasses	110	125	0.8	0.742	
	Hat	54	62	0.778	0.741	
	Jeans	111	130	0.793	0.727	
SSD	Coat	348	1802	0.948	0.843	0.7296
	Dress	178	1118	0.916	0.699	
	Jeans	111	713	0.901	0.691	
	Hat	54	471	0.815	0.688	
	Tie	161	1201	0.851	0.638	
TOOD	Coat	348	798	0.957	0.881	0.8162
	Tie	161	372	0.901	0.848	
	Jeans	111	295	0.928	0.817	
	Hat	54	191	0.87	0.788	
	Dress	178	496	0.916	0.76	
YOLOv3	Coat	348	773	0.92	0.84	0.749
	Tie	161	337	0.87	0.793	
	Glasses	110	338	0.836	0.749	
	Jeans	111	287	0.829	0.686	
	Hat	54	176	0.833	0.677	
YOLOF	Coat	348	773	0.92	0.84	0.714
	Tie	161	337	0.87	0.793	
	Glasses	110	338	0.836	0.749	
	Jeans	111	287	0.829	0.686	
	Hat	54	176	0.833	0.677	

TABLE 4.4: AP of top 5 attire labels for applied object detection methods

The table 4.2.8.3 focuses on the evaluation of state-of-the-art algorithms for object detection on the dataset, encompassing analysis of diverse performance metrics such as precision, recall, and mean average precision, aims to test the algorithms on the dataset and also see whether the dataset is comprehensive enough to be deployed in the real world.

In table 4.2.8.3, Guided anchoring demonstrated excellent recall rates across the top five categories, especially when detecting Coats. The mean average Precision score of 0.785 suggests moderate accuracy across the top five attire categories. This also provides insight into the designing and implementation of the method, highlighting a drawback of relying on an anchor-based approach, which is susceptible to the challenges of high in-class imbalance regarding the size of the detected objects as we did not fine-tune these algorithms to our datasets. As a result, the method may successfully identify instances of coats and ties with high recall rates. The method's robustness in detecting smaller objects like Ties and Hats is evident. The precision of these detections may vary due to inaccuracies in anchor box placement and sizing, leading to fluctuations in average precision metrics. We believe that by fine-tuning this algorithm according to our dataset, the average precision can be improved across all attire labels on our dataset.

Another algorithm, Auto-Assign, With a mean average precision score for the top five attire labels of 0.765, is attributed to the design of this algorithm. The Auto-Assign approach to update detection utilises a mechanism that automatically assigns positive and negative samples during training to optimise detection performance. However, this mechanism might not have been very effective on our dataset. The reason is that attire items exhibit significant variations in appearance, texture, and context, and automatically assigning samples might lead to some optimal training outcomes, resulting in lower Precision levels. One of the algorithms detected glasses as one of the top attires, which suggests that this method May excel in detecting certain attire items. However, we believe that carefully optimising and designing the Auto-Assign architecture to our specific dataset has the potential to achieve higher precision levels and be more consistent across attire categories.

Transformer-based object detectors such as DETR-Deform and DETR stand out as top performers, showcasing exceptional performance across all metrics. DETR-Deform achieves

an outstanding mean for every Precision of 0.8336 in the top five attire classes. The method's strong performance across most attire labels, particularly in detecting Dresses and Coats, shows its adaptability.

Comparing DETR-Deform and DETR with other detectors, such as Guided Anchoring and Auto-Assign, further elucidates their relative performance. Guided Anchoring, while exhibiting commendable recall rates, demonstrates a lower mAP with a score of 0.785, suggesting moderate precision across the top 5 attire categories. Similarly, Auto Assign showcases a slightly lower mAP score of 0.765, indicating slightly lower precision than DETR-Deform and DETR. The attention mechanism and the deformable Transformer layers allow DETR-Deformable to dynamically adjust the sampling location of feature maps and enable the models to adapt to the different variations in attire shapes and sizes, leading to a more accurate and precise detection. We aim to develop a dataset that would enable algorithms to be deployed in real-world scenarios. Transformer-based methods such as DETR-Deform and DETR, which capture structural information and object relationships, perform better than their counterparts.

DDOD, FRCNN, SSD, TOOD, YOLOv3, and YOLOF: These methods show varying performance levels across different attire categories. While some achieve high recall rates for specific attire items, others show more balanced performance across multiple categories. TOOD stands out with its high recall and average precision values for coats and ties. One noteworthy observation is the trade-off between precision and recall in methods like FRCNN, which, despite having a moderate mAP₅ of 0.714, exhibits lower recall, indicating potential issues with missed detections. This suggests that while FRCNN might be accurate when it does detect an object, it is less consistent in finding all instances of an object, which is important in real-world deployment of attire detection applications. YOLOv3, despite its popularity, lags behind with a mAP₅ of 0.749, potentially because of its limitations in detecting smaller or more complex attire objects. The significant performance difference between SSD and DDOD further emphasises the importance of modifications in the architecture of modern deep-learning object detection frameworks.

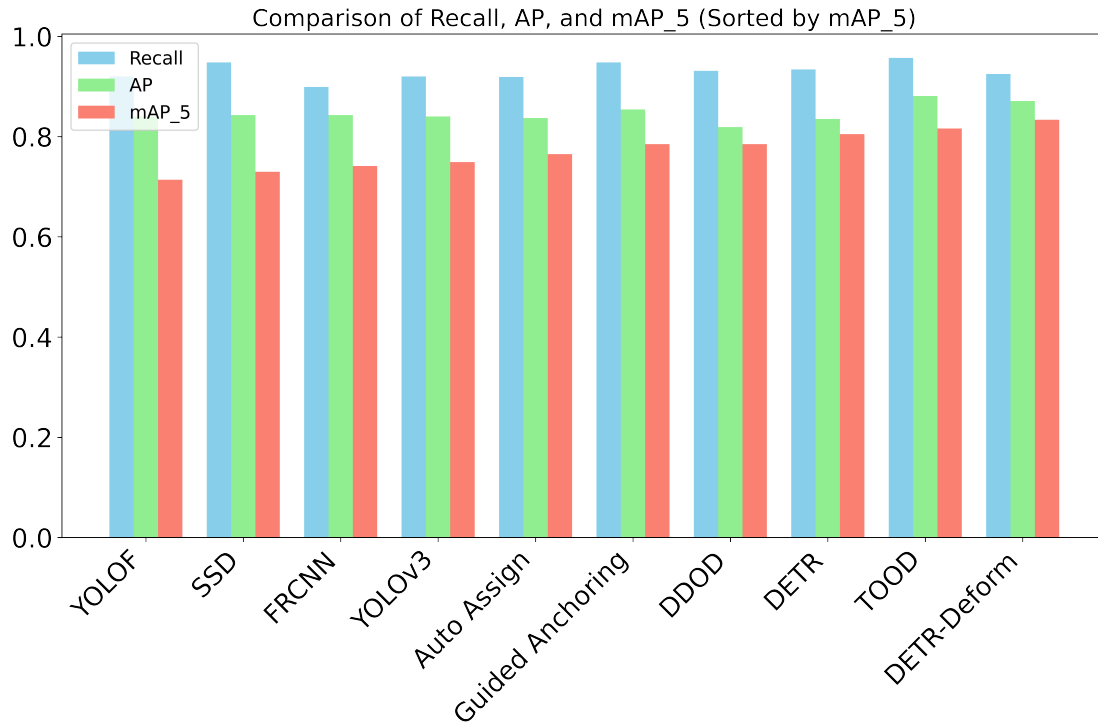


FIGURE 4.12: The chart shows a clear Ranking of Algorithms according to mAP@5. DETR-Deform was the best-performing Algorithm, followed by TOOD.

Transformer-based methods such as DETR-Deform and DETR, which capture structural information and object relationships, perform better than their counterparts. However, further experimentation and fine-tuning may be warranted to optimise detection performance for specific attire items, such as dresses and hats, across all methods.

Researchers can further use the findings from this evaluation to assess the practicality of these object detection algorithms and the promise they hold for real-world applications that require precise estimation from facial images. In conclusion, We believe that this study offers valuable insight into the performance of our dataset in real-world scenarios and sheds light on not only object detection algorithms and their practicality in real-world deployment but also elaborates and examines face detection, alignment, and image enhancement techniques.

4.2.9 Applicability in Real-World Settings



FIGURE 4.13: Snapshots of Implementation System in Mall Setting.

To implement the proposed solution in a mall-like setting, we use a mall video (*Stock Mall Video* n.d.) for detecting faces. We then predict ages, gender, and classify attire of the people in the mall video. The video of the implementation showing the bounding boxes and the predicted age, gender, and attire is shown in (*Mall Video Classification Results* 2024). When it comes to predicting age, the solution works best on the age bracket 20-50 and struggles to correctly identify children because of the training data it has been trained on. A couple of snapshots of the video can be seen in Figure 4.13. In addition to testing the deployment in a

Mall, we also developed an Android App whose APK can be downloaded from (*AndroidApp* 2024).

4.2.10 Heatmaps Providing Insights



FIGURE 4.14: **Top:** The top two images are taken approximately 1 second apart. The left image shows initial predictions, while the right image shows heatmaps to highlight areas where the model focused its attention as the personnel moved, offering insight into the detection process. **Bottom:** Heatmap visualisation of detections in a busy shopping environment. The model's focus on faces and upper bodies is highlighted, with warmer colours indicating higher attention, aiding in understanding detection patterns and potential biases.

Looking at our objective of simultaneous Age, Gender and Attire detection in the wild. I decided to study in depth one of the variations of the YOLO algorithm, you only look one. I choose this algorithm (YOLOv3) because of its effectiveness in real-time object detection

scenarios. Its ability to perform accurate prediction quickly in real time made it suitable for our task of studying in depth how these algorithms would perform in Real-time.

To develop a deeper understanding of how a real-time object detector would make its predictions, I generated heat maps that provided a visual representation of the areas within images and showed where the model concentrated its attention during detection, as seen in Figure 4.14. The heat maps allowed us to analyse and visually inspect the features and regions of different labels that the model deemed most relevant for prediction. This insight was invaluable in confirming that the model was focusing on appropriate and expected regions—such as facial features for age and gender prediction and specific attire elements for clothing classification.

The heat maps also showed an important model behaviour of label bias. Looking at the regions that influence the model's decisions, I was able to identify potential bias in the model's detection process. For example, the model focused its attention on certain attire types, ignoring the environment (context), which resulted in skewed prediction. Identification of such biases informed us about the current limitations of our model and also provides with a direction for future improvement.

In conclusion, the generation of heatmaps provided a powerful combination for both efficient detection and deep interpretative analysis within our thesis. These tools allowed us to critically evaluate and understand the underlying behaviour of our models, setting the stage for further refinement and research in this area.

Future research could explore advanced XAI techniques that go beyond heatmaps, such as feature importance analysis or counterfactual explanations, to further improve model transparency. Furthermore, addressing the label bias identified in our model could involve the development of fairness-aware algorithms or more balanced training datasets.

While our focus has been on real-time detection, future research could investigate how the findings from this study, particularly the use of XAI, can be leveraged to optimise models for low-resource environments. This could involve reducing model complexity without sacrificing accuracy and ensuring that age, gender, and attire detection systems are both effective and

accessible in a wide range of deployment scenarios. For our study, we trained and deployed low-resource variants of YOLO on Android applications.

CHAPTER 5

Conclusion

In the previous chapter, we presented the results of our experiments. In this chapter, we conclude the thesis by summarising our findings and discussing potential avenues for future work.

5.1 Concluding Remarks

The research conducted in this thesis and the newly created dataset has the potential to open new research avenues in computer vision by integrating face, apparent age estimation, and attire detection into a unified framework. The dataset that has been curated has extensive labelling present in the wild, and a variety in the number of faces and the ages of faces. Faces have also been provided in the dataset, cropped and aligned from every image in the dataset, and a dataset with image enhancement through GFP-GAN is also provided. Datasets and their implementation on various state-of-the-art algorithms have demonstrated that some deep learning algorithms can cater to the real-world environment. This is demonstrated through high average precision values for face detection, gender prediction, attire detection, and low mean absolute error values for apparent age estimation. This performance of deep learning algorithms can only be achieved when a carefully annotated dataset is present. Later in this chapter, we also shed light on extending the current findings of this research and exploring further challenges that have arisen during this research.

This thesis has extensively examined face detection, age estimation, attire detection, and related datasets. We have successfully introduced a novel dataset, the first of its kind, that can be used to train deep learning algorithms to simultaneously estimate age, gender, and attire.

This addresses our first research question (**RQ1**), that whether deep-learning architectures can cater for multi-task and multi-class labelling. Our experimental results demonstrate that deep learning algorithms can cater to diverse, multi-labelled, complex data. Furthermore, these algorithms can perform better with hyper-parameter tuning and customise loss functions.

Our analysis demonstrates that certain deep learning algorithms, namely AutoAssign, SSD, FRCNN, DETR, and DETR-Deform exhibit superior performance in apparent age estimation, achieving a mean absolute error of less than 7. Conversely, DETR-Deform, YOLOv3, DDOD, and DETR-Deform are particularly effective for gender prediction, exceeding an average precision of 0.85 for both male and female genders. Finally, in attire detection, DETR, DETR-Deform, and TOOD exhibit a mean average precision of greater than 0.8. Among the algorithms that we analysed, DETR-Deform is the only one that performed effectively for each task. This is because the deformable attention mechanisms in DETR-Deform enables the model to dynamically adjust the spatial sampling locations for attending to features during object detection tasks. This, in turn, improves the model's ability to localise objects accurately in complex scenes. This unique feature contributes to the effectiveness of DETR-Deform in multi-class labelled object detection tasks compared to other deep learning algorithms.

For our second research question (**RQ2**), we observe whether incorporating different percentages of the area around faces impacts the outcomes of deep learning algorithms in predicting age and gender in the wild. Based on our experimental results, we can conclude that incorporating contextual information around the face ($\sim 40\%$ area around the face) helps these deep-learning algorithms in predicting age and gender accurately.

Our last research question (**RQ3**) examines whether image enhancement methods affect age and gender prediction in the wild. We explored the effects of image enhancements on age and gender estimation. We notice that the image quality did not only affect the accuracy of face detection and age estimation, but it was also instrumental in improving the accuracy of gender classification (*see* Figure 4.9). Based on our findings, we recommend that image enhancement techniques should be made a regular part of deep learning-based object detection systems for real-world applications.

5.2 Future Work

Researchers can further explore the optimisation and enhancement of algorithms to enhance performance and efficiency. More efficient deep learning models, such as GAN-based architectures or Transformers with customised loss functions, can be trained and researched to further enhance the understanding of these algorithms.

Researchers can further explore and expand the data set. While very comprehensive, the data set can be changed to include more diverse visual representations, such as clothing from different cultures and traditions. This internship will improve the accuracy and precision of algorithm strain on the dataset, and the algorithms can be deployed globally in real-world environments.

Researchers can focus on refining the algorithms by developing multitasking learning frameworks capable of efficiently handling diverse labels in real-world environments. The dataset can be used as a benchmark to test these algorithms and real-world environments. Another important thing that could be explored is improving the real-time processing of these algorithms by using this dataset to be deployed in dynamic street scenes or surveillance footage that can expand various research avenues.

The role of image processing and enhancement of images to improve the accuracy of detection and classification models was also shown in this research so that researchers can further explore the effect of image enhancement algorithms on object detection algorithms. Researchers can explore various techniques to cater to different lighting conditions, occlusions and complex backgrounds. Researchers can also explore face alignment methods and techniques to compare age estimation and gender classification results in the wild.

One of the major concerns that have risen since the application of this research is the ethical and practical challenges of face detection, agitation gender classification, and attire detection. Future research should focus on ensuring the privacy and security of all people at all times. One avenue of research that could be targeted is addressing biases in these datasets and models to ensure that these algorithms are fair in their judgement across all races. Research into this

topic can also focus on the deployment of these algorithms on edge devices, mobile devices and the cloud. Researchers can look to optimise these models to be deployed on low-power mobile devices. Extensive real-world testing is also required to validate the effectiveness and reliability of these models.

The integration of age, gender, and attire detection holds immense potential for various applications. However, it is crucial to prioritise fairness, transparency, and ethical considerations throughout the process. By expanding datasets, refining algorithms, optimising image pre-processing, and addressing societal implications, we can create robust and responsible computer vision systems for the future. The research carried out in the thesis lays a promising foundation for further exploration in this domain.

Bibliography

- Abbas, Waseem et al. (2024). ‘AI-Driven Precision Clothing Classification: Revolutionizing Online Fashion Retailing with Hybrid Two-Objective Learning’. In: *Information* 15.4, p. 196. DOI: [10.3390/info15040196](https://doi.org/10.3390/info15040196).
- Adams, Rolf and Leanne Bischof (1994). ‘Seeded region growing’. In: *IEEE Transactions on pattern analysis and machine intelligence* 16.6, pp. 641–647. ISSN: 0162-8828.
- Al Hassan, Muneeb (2018). ‘AlexNet - ImageNet Classification with Convolutional Neural Networks’. In: *Neurohive.io*. URL: <https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/>.
- Alhalabi, M. et al. (2021). ‘Sustainable Smart Advertisement Display Using Deep Age and Gender Recognition’. In: *2021 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 33–37. DOI: [10.1109/DASA53625.2021.9682398](https://doi.org/10.1109/DASA53625.2021.9682398).
- AndroidApp (2024). https://drive.google.com/file/d/1S0J7COd-GtY4Mp0vFD9znwgxmQvWE/view?usp=drive_link.
- Arora, Manju, Sahil Sharma and Mohd Asif Khan (2024). ‘Object detection and age & gender estimation using deep learning: An overview’. In: *International Journal of Computing and Artificial Intelligence* 5, pp. 22–29.
- Bansal, Arpit et al. (2021). ‘MetaBalance: High-Performance Neural Networks for Class-Imbalanced Data’. In: *CoRR* abs/2106.09643. arXiv: [2106.09643](https://arxiv.org/abs/2106.09643). URL: <https://arxiv.org/abs/2106.09643>.
- Belongie, Serge, Jitendra Malik and Jan Puzicha (2002). ‘Shape matching and object recognition using shape contexts’. In: *IEEE transactions on pattern analysis and machine intelligence* 24.4, pp. 509–522. ISSN: 0162-8828.

- Bhardwaj, Harshit and Anshuman Sengar (2024). ‘Age and Gender Detection With Real Time Database’. In: *International Journal of Research in Circuits, Devices and Systems* 5, pp. 11–16.
- Bossard, Lukas et al. (2012). ‘Apparel classification with style’. In: *Asian conference on computer vision*. Springer, pp. 321–335.
- (2013). ‘Apparel classification with style’. In: *Computer Vision—ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV 11*. Bossard, Lukas Dantone, Matthias Leistner, Christian Wengert, Christian Quack, Till Van Gool, Luc: Springer, pp. 321–335. ISBN: 3642374468.
- Bradski, Gary (2000). ‘The opencv library’. In: *Dr. Dobb’s Journal: Software Tools for the Professional Programmer* 25.11, pp. 120–123. ISSN: 1044-789X.
- Bradski, Gary and Adrian Kaehler (2000). ‘OpenCV’. In: *Dr. Dobb’s journal of software tools* 3.2.
- Cai, Zhaowei and Nuno Vasconcelos (2018). ‘Cascade r-cnn: Delving into high quality object detection’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Cai, Zhaowei Vasconcelos, Nuno, pp. 6154–6162.
- Cao, Qiong et al. (2018). ‘Vggface2: A dataset for recognising faces across pose and age’. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. Cao, Qiong Shen, Li Xie, Weidi Parkhi, Omkar M Zisserman, Andrew: IEEE, pp. 67–74. ISBN: 1538623358.
- Carion, Nicolas et al. (2020a). ‘End-to-end object detection with transformers’. In: *European conference on computer vision*. Carion, Nicolas Massa, Francisco Synnaeve, Gabriel Usunier, Nicolas Kirillov, Alexander Zagoruyko, Sergey: Springer, pp. 213–229.
- (2020b). ‘End-to-end object detection with transformers’. In: *European conference on computer vision*. Springer, pp. 213–229.
- Chang, Chin-Chou and Ling-Ling Wang (1996). ‘Color texture segmentation for clothing in a computer-aided fashion design system’. In: *Image and vision computing* 14.9, pp. 685–702. ISSN: 0262-8856.

- Chao, Xiaofei et al. (2009). 'A framework for robust feature selection for real-time fashion style recommendation'. In: *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*. ACM, pp. 35–42. ISBN: 1605587583.
- Chen, Bor-Chun, Chu-Song Chen and Winston H Hsu (2014). 'Cross-age reference coding for age-invariant face recognition and retrieval'. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*. Springer, pp. 768–783. ISBN: 3319105981.
- (2015a). 'Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset'. In: *IEEE Transactions on Multimedia* 17.6, pp. 804–815. ISSN: 1520-9210.
- (2015b). 'Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval'. In: *IEEE Transactions on Multimedia* 17.6, pp. 804–815.
- Chen, Hong et al. (2006). 'Composite templates for cloth modeling and sketching'. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 1. Chen, Hong Xu, Zi Jian Liu, Zi Qiang Zhu, Song Chun: IEEE, pp. 943–950. ISBN: 0769525970.
- Chen, Huizhong, Andrew Gallagher and Bernd Girod (2012). 'Describing clothing by semantic attributes'. In: *European conference on computer vision*. Springer, pp. 609–623.
- Chen, Qiang et al. (2015c). 'Deep domain adaptation for describing people based on fine-grained clothing attributes'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Chen, Qiang Huang, Junshi Feris, Rogerio Brown, Lisa M Dong, Jian Yan, Shuicheng, pp. 5315–5324.
- Chen, Yukang et al. (2019). 'Detnas: Neural architecture search on object detection'. In: *arXiv preprint arXiv:1903.10979* 1.2, pp. 4–1.
- Chen, Zehui et al. (2021). 'Disentangle your dense object detector'. In: *Proceedings of the 29th ACM international conference on multimedia*, pp. 4939–4948.
- Chi, Cheng et al. (2019). 'Selective refinement network for high performance face detection'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. Chi, Cheng Zhang, Shifeng Xing, Junliang Lei, Zhen Li, Stan Z Zou, Xudong, pp. 8231–8238. ISBN: 2374-3468.

- Chollet, François (2017). ‘Xception: Deep learning with depthwise separable convolutions’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Chollet, François, pp. 1251–1258.
- Dai, Jifeng et al. (2016). ‘R-fcn: Object detection via region-based fully convolutional networks’. In: *Advances in neural information processing systems* 29.
- Dai, Jifeng et al. (2017). ‘Deformable convolutional networks’. In: *Proceedings of the IEEE international conference on computer vision*. Dai, Jifeng Qi, Haozhi Xiong, Yuwen Li, Yi Zhang, Guodong Hu, Han Wei, Yichen, pp. 764–773.
- Dalal, Navneet and Bill Triggs (2005). ‘Histograms of oriented gradients for human detection’. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Dalal, Navneet Triggs, Bill: Ieee, pp. 886–893. ISBN: 0769523722.
- De Felice, Fabio, Marta Travaglioni and Antonella Petrillo (2023). ‘Innovation Trajectories for a Society 5.0’. In: *Data* 8.10, p. 145. DOI: [10.3390/data8100145](https://doi.org/10.3390/data8100145).
- Deng, Jia et al. (2009). ‘Imagenet: A large-scale hierarchical image database’. In: *2009 IEEE conference on computer vision and pattern recognition*. Deng, Jia Dong, Wei Socher, Richard Li, Li-Jia Li, Kai Fei-Fei, Li: Ieee, pp. 248–255. ISBN: 1424439922.
- Deng, Jiankang et al. (2020). ‘Retinaface: Single-shot multi-level face localisation in the wild’. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Deng, Jiankang Guo, Jia Ververas, Evangelos Kotsia, Irene Zafeiriou, Stefanos, pp. 5203–5212.
- Desai, Chaitanya, Deva Ramanan and Charless C Fowlkes (2011). ‘Discriminative models for multi-class object layout’. In: *International journal of computer vision* 95, pp. 1–12. ISSN: 0920-5691.
- Deshpande, Adit (2016). *A beginner’s guide to understanding Convolutional Neural Networks Part 1*. URL: <https://www.kdnuggets.com/2016/09/beginners-guide-understanding-convolutional-neural-networks-part-1.html>.
- Divvala, Santosh K et al. (2009). ‘An empirical study of context in object detection’. In: *2009 IEEE Conference on computer vision and Pattern Recognition*. Divvala, Santosh K Hoiem, Derek Hays, James H Efron, Alexei A Hebert, Martial: IEEE, pp. 1271–1278. ISBN: 1424439922.

- Duan, Kaiwen et al. (2019). ‘Centernet: Keypoint triplets for object detection’. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578.
- Eidinger, Eran, Roe Enbar and Tal Hassner (2014a). ‘Age and Gender Estimation of Unfiltered Faces’. In: *IEEE Transactions on Information Forensics and Security* 9.12, pp. 2170–2179.
- (2014b). ‘Age and gender estimation of unfiltered faces’. In: *IEEE Transactions on information forensics and security* 9.12, pp. 2170–2179. ISSN: 1556-6013.
- ELKarazle, Khaled, Valliappan Raman and Patrick Then (2022). ‘Facial Age Estimation Using Machine Learning Techniques: An Overview’. In: *Big Data and Cognitive Computing* 6.4, p. 128. DOI: [10.3390/bdcc6040128](https://doi.org/10.3390/bdcc6040128). URL: <https://www.mdpi.com/2504-2289/6/4/128>.
- Erhan, Dumitru et al. (2014). ‘Scalable object detection using deep neural networks’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Erhan, Dumitru Szegedy, Christian Toshev, Alexander Anguelov, Dragomir, pp. 2147–2154.
- Everingham, Mark et al. (2010). ‘The pascal visual object classes (voc) challenge’. In: *International journal of computer vision* 88, pp. 303–338. ISSN: 0920-5691.
- Felzenszwalb, Pedro, David McAllester and Deva Ramanan (2008). ‘A discriminatively trained, multiscale, deformable part model’. In: *2008 IEEE conference on computer vision and pattern recognition*. Felzenszwalb, Pedro McAllester, David Ramanan, Deva: Ieee, pp. 1–8. ISBN: 1424422426.
- Felzenszwalb, Pedro F et al. (2009). ‘Object detection with discriminatively trained part-based models’. In: *IEEE transactions on pattern analysis and machine intelligence* 32.9, pp. 1627–1645. ISSN: 0162-8828.
- Feng, Chengjian et al. (2021a). ‘Tood: Task-aligned one-stage object detection’. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, pp. 3490–3499. ISBN: 1665428120.
- Feng, Yuantao et al. (2021b). ‘Detect faces efficiently: A survey and evaluations’. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.1, pp. 1–18. ISSN: 2637-6407.

- Freixenet, Jordi et al. (2002). ‘Yet another survey on image segmentation: Region and boundary information integration’. In: *European Conference on Computer Vision*. Springer, pp. 408–422.
- Fu, Cheng-Yang et al. (2017). ‘Dssd: Deconvolutional single shot detector’. In: *arXiv preprint arXiv:1701.06659*.
- Gallagher, Andrew C. and Tsuhan Chen (2008). ‘Clothing Cosegmentation for Recognizing People’. In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.
- Ge, Yuying et al. (2019). ‘DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images’. In: *arXiv preprint arXiv:1901.07973*.
- Geirhos, Robert et al. (2020). ‘Shortcut Learning in Deep Neural Networks’. In: *Papers With Code*. URL: <https://paperswithcode.com/paper/shortcut-learning-in-deep-neural-networks>.
- Georgopoulos, Markos, Yannis Panagakis and Maja Pantic (2020). ‘Investigating bias in deep face analysis: The kanface dataset and empirical study’. In: *Image and vision computing* 102, p. 103954. ISSN: 0262-8856.
- Ghani, Arfan et al. (2022). ‘Accelerated Diagnosis of Novel Coronavirus (COVID-19)—Computer Vision with Convolutional Neural Networks (CNNs)’. In: *Electronics* 11.7, p. 1148. DOI: [10.3390/electronics11071148](https://doi.org/10.3390/electronics11071148). URL: <https://www.mdpi.com/2504-2289/6/4/128>.
- Girshick, Ross (2015). ‘Fast r-cnn’. In: *Proceedings of the IEEE international conference on computer vision*. Girshick, Ross, pp. 1440–1448.
- Girshick, Ross et al. (2014). ‘Rich feature hierarchies for accurate object detection and semantic segmentation’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Girshick, Ross Donahue, Jeff Darrell, Trevor Malik, Jitendra, pp. 580–587.
- Gupta, Saurabh, Bharath Hariharan and Jitendra Malik (2015). ‘Exploring person context and local scene context for object detection’. In: *arXiv preprint arXiv:1511.08177*.

- Hadi Kiapour, M et al. (2015). ‘Where to buy it: Matching street clothing photos in on-line shops’. In: *Proceedings of the IEEE international conference on computer vision*, pp. 3343–3351.
- Han, Song, Huizi Mao and William J Dally (2015). ‘Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding’. In: *arXiv preprint arXiv:1510.00149*.
- Hasan, Basela and David C Hogg (2010). ‘Segmentation using Deformable Spatial Priors with Application to Clothing’. In: *BMVC*. Hasan, Basela Hogg, David C, pp. 1–11.
- He, Kaiming and Jian Sun (2015). ‘Convolutional neural networks at constrained time cost’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. He, Kaiming Sun, Jian, pp. 5353–5360.
- He, Kaiming et al. (2015). ‘Spatial pyramid pooling in deep convolutional networks for visual recognition’. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9, pp. 1904–1916. ISSN: 0162-8828.
- (2016). ‘Deep residual learning for image recognition’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. He, Kaiming Zhang, Xiangyu Ren, Shaoqing Sun, Jian, pp. 770–778.
- He, Kaiming et al. (2017). ‘Mask r-cnn’. In: *Proceedings of the IEEE international conference on computer vision*. He, Kaiming Gkioxari, Georgia Dollár, Piotr Girshick, Ross, pp. 2961–2969.
- Hosang, Jan et al. (2015). ‘What makes for effective detection proposals?’ In: *IEEE transactions on pattern analysis and machine intelligence* 38.4, pp. 814–830. ISSN: 0162-8828.
- Hu, Han et al. (2018). ‘Relation networks for object detection’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Hu, Han Gu, Jiayuan Zhang, Zheng Dai, Jifeng Wei, Yichen, pp. 3588–3597.
- Hu, Peiyun and Deva Ramanan (2017). ‘Finding tiny faces’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Hu, Peiyun Ramanan, Deva, pp. 951–959.

- Huang, Huaibo et al. (2017). ‘Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution’. In: *Proceedings of the IEEE international conference on computer vision*, pp. 1689–1697.
- Huang, Junshi et al. (2015). ‘Cross-domain image retrieval with a dual attribute-aware ranking network’. In: *Proceedings of the IEEE international conference on computer vision*, pp. 1062–1070.
- Huang, Zedong et al. (2022). ‘Deep Learning for Computer Vision: Algorithms, Theory and Application’. In: *Electronics* 11.12, p. 1812. DOI: [10.3390/electronics11121812](https://doi.org/10.3390/electronics11121812). URL: <https://www.mdpi.com/2504-2289/6/4/128>.
- Jaccard, Paul (1901). ‘Étude comparative de la distribution florale dans une portion des Alpes et des Jura’. In: *Bull Soc Vaudoise Sci Nat* 37, pp. 547–579.
- Jain, Vidit and Erik Learned-Miller (2010). *Fddb: A benchmark for face detection in unconstrained settings*. Report. UMass Amherst technical report.
- Jiang, Peiyuan et al. (2022). ‘A Review of Yolo algorithm developments’. In: *Procedia Computer Science* 199, pp. 1066–1073. ISSN: 1877-0509.
- Kang, Kai et al. (2017). ‘T-cnn: Tubelets with convolutional neural networks for object detection from videos’. In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.10, pp. 2896–2907. ISSN: 1051-8215.
- Kärkkäinen, Kimmo and Jungseock Joo (2019). ‘Fairface: Face attribute dataset for balanced race, gender, and age’. In: *arXiv preprint arXiv:1908.04913*.
- Karpathy, Andrej and Li Fei-Fei (2015). ‘Deep visual-semantic alignments for generating image descriptions’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Karpathy, Andrej Fei-Fei, Li, pp. 3128–3137.
- King, Davis E (2009). ‘Dlib-ml: A machine learning toolkit’. In: *The Journal of Machine Learning Research* 10, pp. 1755–1758. ISSN: 1532-4435.
- Krasin, Ivan et al. (2017). ‘Openimages: A public dataset for large-scale multi-label and multi-class image classification’. In: *Dataset available from <https://github.com/openimages>* 2.3, p. 18.
- Krizhevsky, Alex and Geoffrey Hinton (2009). ‘Learning multiple layers of features from tiny images’. In.

- Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). ‘Imagenet classification with deep convolutional neural networks’. In: *Advances in neural information processing systems* 25.
- Kuprashevich, Maksim and Irina Tolstykh (2023). ‘MiVOLO: Multi-input Transformer for Age and Gender Estimation’. In: *Papers With Code*. URL: <https://paperswithcode.com/paper/mivolo-multi-input-transformer-for-age-and-gender-estimation>.
- Kuznetsova, Alina et al. (2020). ‘The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale’. In: *International Journal of Computer Vision* 128.7, pp. 1956–1981. ISSN: 0920-5691.
- Lanitis, Andreas, Chris Taylor and Tim Cootes (2002). ‘Toward Automatic Simulation of Aging Effects on Face Images’. In: *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE.
- Lao, Brian and Karthik Jagadeesh (2016). ‘Convolutional neural networks for fashion classification and object detection’. In: *CCCV 2015: Computer Vision*, pp. 120–129.
- Law, Hei and Jia Deng (2018). ‘Cornersnet: Detecting objects as paired keypoints’. In: *Proceedings of the European conference on computer vision (ECCV)*. Law, Hei Deng, Jia, pp. 734–750.
- LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (2015). ‘Deep learning’. In: *nature* 521.7553, pp. 436–444. ISSN: 0028-0836.
- Li, Haoxiang et al. (2015). ‘A convolutional neural network cascade for face detection’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Li, Haoxiang Lin, Zhe Shen, Xiaohui Brandt, Jonathan Hua, Gang, pp. 5325–5334.
- Li, Jian et al. (2019). ‘DSFD: dual shot face detector’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Li, Jian Wang, Yabiao Wang, Changan Tai, Ying Qian, Jianjun Yang, Jian Wang, Chengjie Li, Jilin Huang, Feiyue, pp. 5060–5069.
- Li, Zeming et al. (2017). ‘Light-head r-cnn: In defense of two-stage object detector’. In: *arXiv preprint arXiv:1711.07264*.

- Lim, Bee et al. (2017). ‘Enhanced deep residual networks for single image super-resolution’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144.
- Lin, Tsung-Yi et al. (2014). ‘Microsoft coco: Common objects in context’. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Lin, Tsung-Yi Maire, Michael Belongie, Serge Hays, James Perona, Pietro Ramanan, Deva Dollár, Piotr Zitnick, C Lawrence: Springer, pp. 740–755. ISBN: 3319106015.
- Lin, Tsung-Yi et al. (2017a). ‘Feature pyramid networks for object detection’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Lin, Tsung-Yi Dollár, Piotr Girshick, Ross He, Kaiming Hariharan, Bharath Belongie, Serge, pp. 2117–2125.
- Lin, Tsung-Yi et al. (2017b). ‘Focal loss for dense object detection’. In: *Proceedings of the IEEE international conference on computer vision*. Lin, Tsung-Yi Goyal, Priya Girshick, Ross He, Kaiming Dollár, Piotr, pp. 2980–2988.
- Lin, Yiming et al. (2022). ‘Fp-age: Leveraging face parsing attention for facial age estimation in the wild’. In: *IEEE Transactions on Image Processing*. ISSN: 1057-7149.
- Liu, Hongxin, Xiaorong Shen and Haibing Ren (2016a). ‘FDAR-Net: Joint convolutional neural networks for face detection and attribute recognition’. In: *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*. Vol. 2. Liu, Hongxin Shen, Xiaorong Ren, Haibing: IEEE, pp. 184–187. ISBN: 1509035583.
- Liu, Kuan-Hsien, Ting-Yen Chen and Chu-Song Chen (2016b). ‘Mvc: A dataset for view-invariant clothing retrieval and attribute prediction’. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, pp. 313–316. ISBN: 1450343597.
- Liu, Songtao and Di Huang (2018). ‘Receptive field block net for accurate and fast object detection’. In: *Proceedings of the European conference on computer vision (ECCV)*. Liu, Songtao Huang, Di, pp. 385–400.
- Liu, Wei et al. (2016c). ‘Ssd: Single shot multibox detector’. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Liu, Wei Anguelov, Dragomir Erhan, Dumitru Szegedy, Christian Reed, Scott Fu, Cheng-Yang Berg, Alexander C: Springer, pp. 21–37. ISBN: 3319464477.

- Liu, Yong et al. (2018). ‘Structure inference net: Object detection using scene-level context and instance-level relationships’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Liu, Yong Wang, Ruiping Shan, Shiguang Chen, Xilin, pp. 6985–6994.
- Liu, Ziwei et al. (2016d). ‘Deepfashion: Powering robust clothes recognition and retrieval with rich annotations’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104.
- Lowe, David G (1999). ‘Object recognition from local scale-invariant features’. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Lowe, David G: Ieee, pp. 1150–1157. ISBN: 0769501648.
- Malisiewicz, Tomasz, Abhinav Gupta and Alexei A Efros (2011). ‘Ensemble of exemplar-svm for object detection and beyond’. In: *2011 International conference on computer vision*. Malisiewicz, Tomasz Gupta, Abhinav Efros, Alexei A: IEEE, pp. 89–96. ISBN: 1457711028.
- Mall Video Classification Results* (2024). <https://drive.google.com/file/d/1o17-7YytXRUOMKgoig5PHW06hZIvicBn/view>.
- Mohammed, Ahmed, Hisham Ahmed and Muhammad Khan (2022). ‘Deep Learning Methods for Real-Time Image-Based Gender Classification in Security Systems’. In: *Computational Intelligence and Neuroscience 2022*, pp. 1–12. DOI: [10.1155/2022/1503188](https://doi.org/10.1155/2022/1503188). URL: <https://www.hindawi.com/journals/cin/2022/1503188/>.
- Moschoglou, Stylianos et al. (2017). ‘Agedb: the first manually collected, in-the-wild age database’. In: *proceedings of the IEEE conference on computer vision and pattern recognition workshops*. Moschoglou, Stylianos Papaioannou, Athanasios Sagonas, Christos Deng, Jiankang Kotsia, Irene Zafeiriou, Stefanos, pp. 51–59.
- Nada, Hajime et al. (2018). ‘Pushing the limits of unconstrained face detection: a challenge dataset and baseline results’. In: *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. Nada, Hajime Sindagi, Vishwanath A Zhang, He Patel, Vishal M: IEEE, pp. 1–10. ISBN: 1538671808.

- Najibi, Mahyar et al. (2017). ‘Ssh: Single stage headless face detector’. In: *Proceedings of the IEEE international conference on computer vision*. Najibi, Mahyar Samangouei, Pouya Chellappa, Rama Davis, Larry S, pp. 4875–4884.
- Ng, Hing N and Richard L Grimsdale (1996). ‘Computer graphics techniques for modeling cloth’. In: *IEEE Computer Graphics and Applications* 16.5, pp. 28–41. ISSN: 0272-1716.
- Özbulak, Gökhan, Yusuf Aytar and Hazım Kemal Ekenel (2016). ‘How Transferable are CNN-based Features for Age and Gender Classification?’ In: *arXiv preprint arXiv:1610.00134*.
- Packer, Charles, Julian McAuley and Arnau Ramisa (2018). ‘Visually-aware personalized recommendation using interpretable image representations’. In: *arXiv preprint arXiv:1806.09820*.
- Padilla, Rafael, Sergio L Netto and Eduardo AB Da Silva (2020). ‘A survey on performance metrics for object-detection algorithms’. In: *2020 international conference on systems, signals and image processing (IWSSIP)*. IEEE, pp. 237–242. ISBN: 1728175399.
- Pandey, Osha (2023). *Age and Gender Detection using Deep Learning*. URL: https://github.com/OshaPandey/Age_Gender_Detection.
- Ponce-López, Víctor et al. (2016). ‘Chalearn lap 2016: First round challenge on first impressions-dataset and results’. In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, pp. 400–418. ISBN: 3319494082.
- Pramilarani, K. et al. (2024). ‘Empowering Law Enforcement and Emergency: An In-Depth Exploration of Criminal Face Detection Technology Integrated with Emergency SOS Sending Protocols’. In: *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*. Bangalore, India, pp. 1–6. DOI: [10.1109/ICDECS59733.2023.10503524](https://doi.org/10.1109/ICDECS59733.2023.10503524).
- Redmon, Joseph et al. (2016). ‘You only look once: Unified, real-time object detection’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Redmon, Joseph Divvala, Santosh Girshick, Ross Farhadi, Ali, pp. 779–788.
- Ren, Shaoqing et al. (2015). ‘Faster r-cnn: Towards real-time object detection with region proposal networks’. In: *Advances in neural information processing systems* 28.
- Ricanek, Karl and Tamirat Tesafaye (2006a). ‘Morph: A longitudinal image database of normal adult age-progression’. In: *7th international conference on automatic face and*

- gesture recognition (FGR06)*. Ricanek, Karl Tesafaye, Tamirat: IEEE, pp. 341–345. ISBN: 0769525032.
- (2006b). ‘Morph: A longitudinal image database of normal adult age-progression’. In: *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, pp. 341–345. ISBN: 0769525032.
- Ricanek, Karl Jr. and Tamirat Tesafaye (2006c). ‘MORPH: A Longitudinal Image Database of Normal Adult Age-Progression’. In: *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*.
- Rothe, Rasmus, Radu Timofte and Luc Van Gool (2016). ‘Do We Really Need to Collect Millions of Faces for Effective Face Recognition?’ In: *European Conference on Computer Vision*. Springer, pp. 579–596.
- Rothe, Rasmus, Radu Timofte and Luc Van Gool (2015). ‘Dex: Deep expectation of apparent age from a single image’. In: *Proceedings of the IEEE international conference on computer vision workshops*. Rothe, Rasmus Timofte, Radu Van Gool, Luc, pp. 10–15.
- Russell, Bryan C et al. (2008). ‘LabelMe: a database and web-based tool for image annotation’. In: *International journal of computer vision* 77, pp. 157–173. ISSN: 0920-5691.
- Siddiqi, Muhammad Hameed et al. (2022). ‘Face Image Analysis Using Machine Learning: A Survey on Recent Trends and Applications’. In: *Electronics* 11.8, p. 1210. DOI: [10.3390/electronics11081210](https://doi.org/10.3390/electronics11081210). URL: <https://www.mdpi.com/2504-2289/6/4/128>.
- Somanath, Gowri, Rohith MV and Chandra Kambhamettu (2011). *VADANA: Vims Appearance Dataset for Facial Image ANALysis*. URL: <https://www.sciweavers.org/datasets/vadana-vims-appearance-dataset-facial-image-analysis>.
- Soo, Sander (2014). ‘Object detection using Haar-cascade Classifier’. In: *Institute of Computer Science, University of Tartu* 2.3, pp. 1–12.
- Stock Mall Video* (n.d.). <https://www.pexels.com/video/people-460walking-inside-the-shopping-mall-4750056>.
- Szegedy, Christian et al. (2017). ‘Inception-v4, inception-resnet and the impact of residual connections on learning’. In: *Proceedings of the AAAI conference on artificial intelligence*.

- Vol. 31. Szegedy, Christian Ioffe, Sergey Vanhoucke, Vincent Alemi, Alexander. ISBN: 2374-3468.
- Tang, Xu et al. (2018). ‘Pyramidbox: A context-assisted single shot face detector’. In: *Proceedings of the European conference on computer vision (ECCV)*. Tang, Xu Du, Daniel K He, Zeqiang Liu, Jingtuo, pp. 797–813.
- Tian, Zhi et al. (2019). ‘Fcos: Fully convolutional one-stage object detection’. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Tian, Zhi Shen, Chunhua Chen, Hao He, Tong, pp. 9627–9636.
- Twigg, Julia (2007). ‘Clothing, age and the body: A critical review’. In: *Ageing & Society* 27.2, pp. 285–305. ISSN: 0144-686X.
- Tzutalin (2015). *labelImg: LabelImg is a graphical image annotation tool and label object bounding boxes in images*. <https://github.com/HumanSignal/labelImg>.
- Uijlings, Jasper RR et al. (2013). ‘Selective search for object recognition’. In: *International journal of computer vision* 104, pp. 154–171. ISSN: 0920-5691.
- Vaswani, Ashish et al. (2017). ‘Attention is all you need’. In: *Advances in neural information processing systems* 30.
- Viola, Paul and Michael Jones (2001). ‘Rapid object detection using a boosted cascade of simple features’. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Viola, Paul Jones, Michael: Ieee, pp. I–I. ISBN: 0769512720.
- Wang, Chien-Yao, Alexey Bochkovskiy and Hong-Yuan Mark Liao (2023). ‘YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Wang, Chien-Yao Bochkovskiy, Alexey Liao, Hong-Yuan Mark, pp. 7464–7475.
- Wang, Hao et al. (2017). ‘Face r-cnn’. In: *arXiv preprint arXiv:1706.01061*.
- Wang, Nan and Haizhou Ai (2011). ‘Who blocks who: Simultaneous clothing segmentation for grouping images’. In: *2011 International Conference on Computer Vision*. Wang, Nan Ai, Haizhou: IEEE, pp. 1535–1542. ISBN: 1457711028.

- Wang, Xintao et al. (2018). 'Esrgan: Enhanced super-resolution generative adversarial networks'. In: *Proceedings of the European Conference on computer vision (ECCV) workshops*, pp. 0–0.
- Wang, Xintao et al. (2021). 'Towards real-world blind face restoration with generative facial prior'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9168–9178.
- Xiang, Jia and Gengming Zhu (2017). 'Joint face detection and facial expression recognition with MTCNN'. In: *2017 4th international conference on information science and control engineering (ICISCE)*. IEEE, pp. 424–427. ISBN: 1538630133.
- Yamaguchi, Kota, M Hadi Kiapour and Tamara L Berg (2013). 'Paper doll parsing: Retrieving similar styles to parse clothing items'. In: *Proceedings of the IEEE international conference on computer vision*. Yamaguchi, Kota Hadi Kiapour, M Berg, Tamara L, pp. 3519–3526.
- Yan, Junjie et al. (2014). 'Face detection by structural models'. In: *Image and Vision Computing* 32.10, pp. 790–799. ISSN: 0262-8856.
- Yang, Bin et al. (2015). 'Fine-grained evaluation on face detection in the wild'. In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. Yang, Bin Yan, Junjie Lei, Zhen Li, Stan Z: IEEE, pp. 1–7. ISBN: 1479960268.
- Yang, Jun et al. (2007). 'Evaluating bag-of-visual-words representations in scene classification'. In: *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, pp. 197–206. ISBN: 1595937781.
- Yang, Ming and Kai Yu (2011). 'Real-time clothing recognition in surveillance videos'. In: *2011 18th IEEE international conference on image processing*. Yang, Ming Yu, Kai: IEEE, pp. 2937–2940. ISBN: 1457713039.
- Yang, Shuo et al. (2016). 'Wider face: A face detection benchmark'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Yang, Shuo Luo, Ping Loy, Chen-Change Tang, Xiaoou, pp. 5525–5533.

- Yang, Ze et al. (2019). ‘Reppoints: Point set representation for object detection’. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Yang, Ze Liu, Shaohui Hu, Han Wang, Liwei Lin, Stephen, pp. 9657–9666.
- Yoo, YoungJoon, Dongyoon Han and Sangdoo Yun (2019). ‘Extd: Extremely tiny face detector via iterative filter reuse’. In: *arXiv preprint arXiv:1906.06579*.
- Yousuf, Saad Bin et al. (2019). ‘IMDB-attire: A novel dataset for Attire Detection and localization’. In: *Neural Information Processing*, pp. 543–553. DOI: [10.1007/978-3-030-36711-4_46](https://doi.org/10.1007/978-3-030-36711-4_46).
- Zeiler, Matthew D and Rob Fergus (2014). ‘Visualizing and understanding convolutional networks’. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*. Zeiler, Matthew D Fergus, Rob: Springer, pp. 818–833. ISBN: 3319105892.
- Zhang, Kaipeng et al. (2016). ‘Joint face detection and alignment using multitask cascaded convolutional networks’. In: *IEEE signal processing letters* 23.10, pp. 1499–1503. ISSN: 1070-9908.
- Zhang, Shifeng et al. (2017a). ‘S3fd: Single shot scale-invariant face detector’. In: *Proceedings of the IEEE international conference on computer vision*. Zhang, Shifeng Zhu, Xiangyu Lei, Zhen Shi, Hailin Wang, Xiaobo Li, Stan Z, pp. 192–201.
- Zhang, Shifeng et al. (2020). ‘Refineface: Refinement neural network for high performance face detection’. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11, pp. 4008–4020. ISSN: 0162-8828.
- Zhang, Wei et al. (2008). ‘An intelligent fitting room using multi-camera perception’. In: *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, pp. 60–69. ISBN: 1595939873.
- Zhang, Xiangyu et al. (2015). ‘Accelerating very deep convolutional networks for classification and detection’. In: *IEEE transactions on pattern analysis and machine intelligence* 38.10, pp. 1943–1955. ISSN: 0162-8828.
- Zhang, Xishan et al. (2017b). ‘Trip outfits advisor: Location-oriented clothing recommendation’. In: *IEEE Transactions on Multimedia* 19.11, pp. 2533–2544. ISSN: 1520-9210.

- Zhang, Yunxuan, Li Liu and Cheng Li (2017c). ‘Quantifying facial age by posterior of age comparisons’. In: *arXiv preprint arXiv:1708.09687*.
- Zhang, Zhifei, Yang Song and Hairong Qi (2017d). ‘Age progression/regression by conditional adversarial autoencoder’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Zhang, Zhifei Song, Yang Qi, Hairong, pp. 5810–5818.
- (2017e). ‘Age progression/regression by conditional adversarial autoencoder’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818.
- Zhou, Xingyi, Dequan Wang and Philipp Krähenbühl (2019). ‘Objects as points’. In: *arXiv preprint arXiv:1904.07850*.
- Zhu, Benjin et al. (2020a). ‘Autoassign: Differentiable label assignment for dense object detection’. In: *arXiv preprint arXiv:2007.03496*.
- Zhu, Chenchen et al. (2017). ‘Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection’. In: *Deep learning for biometrics*, pp. 57–79. ISSN: 3319616560.
- Zhu, Jiashu et al. (2020b). ‘Progressface: Scale-aware progressive learning for face detection’. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Zhu, Jiashu Li, Dong Han, Tiantian Tian, Lu Shan, Yi: Springer, pp. 344–360. ISBN: 3030585387.
- Zhu, Xiangxin and Deva Ramanan (2012). ‘Face detection, pose estimation, and landmark localization in the wild’. In: *2012 IEEE conference on computer vision and pattern recognition*. Zhu, Xiangxin Ramanan, Deva: IEEE, pp. 2879–2886. ISBN: 1467312282.
- Zhu, Xizhou et al. (2020c). ‘Deformable detr: Deformable transformers for end-to-end object detection’. In: *arXiv preprint arXiv:2010.04159*.
- Zou, Zhengxia et al. (2023). ‘Object detection in 20 years: A survey’. In: *Proceedings of the IEEE*. ISSN: 0018-9219.