Deep Learning Algorithm for Virtual Advertisement Replacement on Fence Board of Sports Games Videos

ZIJUN(KEANE) LI

A thesis submitted to the Auckland University of Technology

in partial fulfilment of the requirements for the degree of

Master of Computer and Information Sciences (MCIS)

2023

School of Engineering, Computer & Mathematical Sciences

Abstract

In the rapidly evolving field of sports broadcasting, the demand for effective and dynamic advertisement placement has increased dramatically. This thesis presents a deep learning-based solution for virtual advertisement replacement on fence boards in sports videos. Our approach leverages Computer Vision (CV) techniques to provide automatic and precise ad placement, even in the presence of common obstacles such as occlusion.

The methodology consists of three main stages. In the first stage, we employ a Key-point Region-Based Convolutional Neural Network (Key-point RCNN) model for fence board detection in the video frames, which is trained and fine-tuned using a manually labelled dataset. This enables the model to identify the optimal locations for placing advertisements (Ads). However, object occlusion, especially by humans, could affect the Ads visibility and effectiveness. We introduce a solution in the second stage using a Matting Objective Decomposition Network (MODNet) model to handle this challenge. It is used for segmenting the human element from the video frames and re-applying them after the Ads are placed on the fence boards. In the final stage, FlowNet is utilized to monitor the placement of Ads. This ensures that when the scene transitions, the new scene is promptly detected, and ad placement is deferred until the fence board or the optimal location, becomes visible.

Our results indicate that this approach achieves high accuracy in fence board detection and human occlusion handling, overcoming the limitations of previous techniques. The proposed solution opens new possibilities in virtual advertisement replacement in sports videos, presenting a promising avenue for further research and practical applications.

Table of Contents

List of Figure

List of Acronyms

Ads ..... Advertisements

AI ..... Artificial Intelligence

AR ..... Augmented Reality

ASPP ..... Atrous Spatial Pyramid Pooling

BGR ..... Blue-Green-Red

CNN ..... Convolutional Neural Network

CV ..... Computer Vision

DL ..... Deep Learning

FPN ..... Feature Pyramid Network

GANs ..... Generative Adversarial Networks

Key-point RCNN ..... Key-point Region-Based Convolutional Neural Network

Key-point RCNN-PF ..... Key-point RCNN - Picture Frame

ML ..... Machine Learning

MODNet ..... Matting Objective Decomposition Network

PF branch ..... Picture Frame branch

RGB ..... Red-Green-Blue

ROI ..... Region of Interest

RPN ..... Region Proposal Network

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 17 July 2023

Acknowledgment

First and foremost, I would like to express my sincere gratitude to everyone who supported and assisted me throughout the course of this research, particularly my parents and family. Their selfless help and generous funding have provided me with the invaluable opportunity to complete my master's studies at the Auckland University of Technology (AUT) in New Zealand.

I also wish to convey my profound appreciation to my primary supervisor, Peter Chong. His professional knowledge, meticulous guidance, and efforts to enrich my learning experience have been instrumental throughout this process.

Lastly, my thanks go to the school administrators for their continual support during my time at AUT. Their backing has been essential to my success.

Zijun Li

Auckland, New Zealand

July 2023

Chapter1 Introduction

1.1 Background and Motivation

The previous decade has been marked by an accelerating interplay of technology within sports broadcasting, especially the incorporation of virtual advertisement placements in sports videos. This technological advent has resulted in a tectonic shift from traditional advertising paradigms, paving the way for inventive and dynamic advertising strategies that boost viewer engagement. Such transformational changes accentuate the necessity of comprehensive research into the multifaceted area of virtual advertisement placement in sports videos.

Virtual advertisement placement in sports videos is an expanding field of study investigating the incorporation of Ads into sports broadcasts. Several subfields have been identified within this domain, such as the effects of virtual advertising, automated content placement, sports video summarisation, the role of social media in advertising, background detection, advertisement positioning, and human occlusion. The efficacy of virtual advertising in sports events has been extensively studied, focusing on the contributing factors to its success. Research indicates that participants highly recognize virtual advertising, thus reinforcing its effectiveness (Sander & Altobelli, 2010; Sander & Altobelli, 2011). Numerous elements, including conspicuousness, brand awareness, and frequency of exposure, are influential in shaping the impact of virtual advertising (Tellis, 1997; Longman, 1997; Pieters & Bijmolt, 1997; Turley & Shannon, 2000), and attitudes towards advertising have also been analyzed as part of these studies, contributing to our understanding of the mechanisms that drive the success of virtual advertising in sports events (Sander & Altobelli, 2011). The financial investment in virtual advertising underscores its potential value. In 2004, for instance, it was estimated that over $2 billion was invested in virtual advertising in the U.S. alone (Pyun et al., 2004), underlining the significant potential of this form of advertising.

By strategically integrating Ads into video content, automated content placement in sports highlights may increase viewer engagement (Wan & Xu, 2006). Vasudevan and Gounder (2021)'s research on sports video summarisation has centred on techniques for detecting highlights, creating customised summaries, and analysing subtitle content to increase content accessibility and engagement. Advances in sports video summarization techniques, such as those implemented in cricket videos, provide insights into customised highlight generation and content analysis. Furthermore, social media's role as a potent advertising platform within the

sports industry warrants investigation due to its expansive reach and influence (Byon & Phua, 2021).

Even though prior research has made significant contributions to the domain, there are several limitations and knowledge gaps to consider. Beyond the immediate effects studied to date, virtual Ads long-term effects and efficacy require additional research (Sander & Altobelli, 2011). Although methods for automated content placement have been proposed, optimising algorithms and evaluating their performance in real-world scenarios remain essential research objectives (Wan & Xu, 2006). Furthermore, while standardising approaches to sports video summarization can enhance content accessibility and viewer engagement, a persistent need for more extensive and comprehensive strategies exists (Vasudevan & Gounder, 2021).

Concurrently, the advent of advanced technologies has unveiled new exploration avenues. Researchers have explored various aspects of this field, including interactive advertising ((Viktoratos & Tsadiras, 2021; Giombi et al., 2022), augmented reality (AR) (Du et al., 2022; Carmigniani et al., 2011), and personalized computational techniques (Viktoratos & Tsadiras, 2021). Virtual Ads can increase consumer engagement and comprehension by integrating interactive elements like gamification or personalised interactions. AR technologies have emerged as a promising instrument for replacing virtual Ads. AR improves the user experience by superimposing digital content on real-world objects, creating immersive and interactive advertising experiences (Carmigniani et al., 2011; Du et al., 2022). Similarly, personalised computational methods play a crucial role in virtual advertisement replacement. Personalised Ads can be tailored to an individual's preferences using artificial intelligence (AI) algorithms and semantic web technologies, increasing their effectiveness (Morrison-Smith & Ruiz, 2020; Viktoratos & Tsadiras, 2021). Additionally, as Li et al. (2005) suggest, future research should concentrate on real-time advertisement insertion and evaluating the effects of Ads in specific sports contexts. Also, investigating the efficacy of various virtual advertisement positioning techniques can contribute to developing advertising guidelines and best practices.

While these existing techniques contribute significantly to the field, additional research is required to fill in several voids. The swift pace of technological progress underscores the importance of exploring newer methods of virtual advertisement replacement, ensuring that the techniques used keep pace with the ever-evolving technological landscape. Given the accelerated advancements in computer vision (CV) and deep learning (DL), it is essential to continue researching the replacement of virtual Ads within this domain.

Various methodologies and models have emerged in the rapidly evolving world of CV and DL, offering tools to tackle many challenges. In the field of virtual advertisement replacement in sports videos, the use of such models is increasingly relevant and promising. In response to this

context, the research in this thesis focuses on the direction of algorithms and techniques within the CV and DL field. It concentrates on proposing a CV-based solution to replace virtual Ads on the fence boards of sports game videos. The aim is to advance the field further and provide a practical solution for enhancing virtual advertisement placement in sports videos.

For the task of virtual ad replacement in sports videos, this thesis focuses on three pivotal models. The Key-point Region-Based Convolutional Neural Network (Key-point RCNN-PF) model, a pioneering concept introduced by Zheng et al. (2021), applies a unique methodology for detecting picture frames of objects within images, associating each instance with its respective category labels. This provides a pathway for identifying the background of sports videos, thus enabling targeted advertisement placement on specific items such as fence boards.

The FlowNet model, an ingenious paradigm shift in optical flow estimation, is Enhancing the application scope of these techniques, proposed by Dosovitskiy et al. (2015). FlowNet's convolutional neural network architecture facilitates the learning of optical flow directly from data. This enables tracking object movements across video frames for advertisement replacement, offering a viable solution for ad position tracking.

Lastly, Ke et al. (2022) proposed the Matting Objective Decomposition Network (MODNet) model, a trailblazing shift in portrait matting. With its ability to perform real-time matting from a solitary input image, MODNet assists in identifying and isolating players or subjects within frames. This provides a solution to the problem of human occlusion, thereby allowing for advertisement placement strategies that do not disturb the continuity of the original video.

In summary, this research proposes a CV-based integration model as a solution for the virtual advertisement replacement on the fence board of the sports game’s videos, which was realized through a step-by-step exploration and analysis of three distinctive models: the Key-point RCNN-PF model, the MODNet model, and the FlowNet model.

1.2 Problem Statement

Despite gaining an initial understanding of virtual Ads replacement in sports videos and developing a preliminary solution, several formidable challenges persist in executing virtual ad replacement within this context:

- Detection and localization of the fence board in sports videos: This is the highest priority, as the fence board is the platform for advertisement placement in this study.
- Improvement of detection precision: It is essential to secure a high and acceptable range of detection accuracy for the fence board. This ensures the presentation of the advertisement in a natural, consistent, and uninterrupted manner.
- Insertion of a virtual advertisement and tracking its position: Advertisement insertion and ad position tracking constitute the crux of the subsequent phase. This aspect is closely tied to the accuracy of ad placement in the intended position and the maintenance of the video's consistency and continuity.
- Handling of scene changes: The dynamic nature of sports videos, characterized by frequent scene changes, poses an additional challenge. Hence, a mechanism capable of detecting scene changes promptly and appropriately delaying advertisement delivery is required.
- Resolving human occlusion post-virtual advertisement application: Finally, addressing the occlusion problem induced by human factors in the foreground post-advertisement application is essential. This ensures that the advertisement does not obstruct the visibility of the video content.

1.3 Objectives and Contributions

This thesis aims to develop a robust CV-based model for virtual ad replacement on the fence board of sports videos. This model explicitly addresses challenges associated with the implementation of virtual ad replacement, including fence board detection, ad placement, scene change detection, and the handling of human occlusion. The proposed model will employ a combination of CV techniques and DL algorithms. Specifically, it will integrate Key-point RCNN-PF for the precise detection of fence boards, MODNet for managing human occlusion issues, and FlowNet for the tracking of ad locations and scene changes issues.

This thesis primarily contributes in two ways. Firstly, it proposes an integrated model for virtual advertisement replacement on the fence boards within sports game videos. Secondly, it provides a viable solution and subsequent direction to address the challenges of fence board detection, advertisement insertion and tracking, scene change recognition, and human occlusion handling.

1.4 Thesis Organization

The structure of this thesis is as follows: Chapter 2 presents the literature review, discussing the evolution and applications of CV and DL in advertisement replacement, an overview of Convolutional Neural Network (CNN), key-point detection, optical flow measurement (or background tracking), and human segmentation and matting.

Chapter 3 explains the methodological theory related to this study, which includes the detection of fence boards using Key-point RCNN-PF, ad location tracking with FlowNet and human occlusion handling with MODNet. It will also elaborate on how the initial integration model functions.

Chapter 4 delves into the respective training, results, and analysis of the above three models. This involves the application of Key-point RCNN-PF for fence board detection, MODNet for human occlusion resolution, and the integration of these models. It will further highlight the role of FlowNet in ad location tracking and detail the results and analysis of applying the initial integration model to a demo video. It will then present the performance and analysis of the final integration model, improved based on the limitations of the initial model, and demonstrate its application in new use cases. Finally, Chapter 5 will conclude the thesis and propose areas for future research.

Chapter 2 Literature Review

2.1 Overview and Applications of Computer Vision

Computer vision (CV) is an interdisciplinary field that focuses on developing algorithms and techniques to enable computers to gain a high-level understanding of visual data, such as images and videos. It involves extracting meaningful information, recognizing patterns, and making inferences from visual inputs. Over the years, CV has witnessed significant advancements driven by emerging techniques and novel applications (Chai et al., 2021).

CV utilizes algorithms and models to simulate human visual perception and interpret visual data. It encompasses several key techniques, including deep learning (DL) (Voulodimos et al., 2018), generative adversarial networks (GANs) (Aggarwal et al., 2021), and syntactic pattern recognition (Astolfi et al., 2021). Deep learning, particularly Convolutional Neural Networks (CNN), has revolutionized CV by outperforming traditional machine learning (ML) techniques in tasks such as object detection, face recognition, and activity recognition (Chai et al., 2021). GANs have made significant strides in image generation, image-to-image translation, and facial attribute manipulation (Aggarwal et al., 2021). Syntactic pattern recognition methods have been widely employed for CV tasks, offering interpretable and hierarchical representations of patterns in images and videos (Astolfi et al., 2021).

CV finds applications across various domains and industries. In the field of healthcare, it has been utilized for medical image analysis, disease diagnosis, and surgical guidance. For instance, CV techniques have been employed in the automated detection and segmentation of tumours in medical images, leading to improved diagnostic accuracy (Esteva et al., 2021). In the fashion industry, CV enables fashion detection, attribute recognition, and outfit recommendation, enhancing the shopping experience and enabling personalized fashion choices (Cheng et al., 2021). Construction and asset management have also benefited from CV, enabling visual data processing for activity monitoring, safety enhancement, and quality control (Rehman et al., 2022).

2.2 Evolution of Deep Learning Algorithms in Computer Vision

In the late 1980s, the concept of neural networks became widely prevalent in the areas of ML and artificial intelligence (AI), leading to the introduction of innovative methods and network structures. Multilayer Perceptron networks, trained by "Backpropagation" type algorithms, were one of the many innovative methods born out of this period (Sarker, 2021). These networks were one of the first types of artificial neural networks capable of modelling non-linear and complex relationships.

DL, a sub-field of ML, was introduced in 2006 by Hinton, Bengio, and LeCun. This method leveraged the concept of artificial neural networks, using multiple layers to extract high-level features from raw input data (Sarker, 2021; Voulodimos et al., 2018). Following this, several DL algorithms were employed to address various artificial intelligence problems. Notably, CNNs became one of the most significant schemes used in CV tasks. Introduced around the 2010s, CNNs consisted of convolutional, pooling, and fully connected layers and demonstrated remarkable success in the image and video recognition tasks (Voulodimos et al., 2018).

Parallelly, other DL structures like Deep Boltzmann Machines (DBMs) (Hinton & Salakhutdinov, 2012), Deep Belief Networks (DBNs)(Mohamed et al., 2012), and Stacked Denoising Autoencoders (SDAs)(Kandaswamy et al., 2014) were developed and applied in CV problems. DBMs are a type of stochastic neural network, DBNs are generative models with multiple layers of hidden variables, and SDAs, an extension of simple autoencoders, were designed to reconstruct corrupted versions of the input, enabling the extraction of more robust and useful features (Voulodimos et al., 2018).

Over the years, specific CNN models like GoogLeNet, VGG Net, and AlexNet have been widely used in various CV applications, including image classification, object detection, and face detection (Kalaiarasi & Rani, 2020). These models, despite their differences in structure and complexity, are all based on the fundamental principles of convolutional neural networks.

2.3 Computer Vision (CV) and Deep Learning (DL) in Advertisement Replacement

Voulodimos et al. (2018) underscore the substantial advancements made in the fields of CV and DL in recent years, which have found applications in a myriad of domains. DL, in particular, has been instrumental in addressing numerous challenging problems such as CV, image recognition, and image detection, thereby proving its value in various real-world applications. The domain of advertisement replacement has witnessed significant advancements. For instance, Hossari et al. (2018) proposed ADNet, a deep-learning architecture that automatically detects the presence of Ads in video frames. Similarly, Efimova et al. (2022) introduced a novel algorithm for processing and replacing advertising banners in videos, preserving the quality of the original video content. Liu (2023) further contributed by presenting a non-intrusive framework capable of detecting an ad section and segmenting individual ads. This evolution in the field demonstrates a concerted research focus on detecting and recognizing Ads in multimedia data (Almgren et al., 2018).

DL techniques have outpaced their ML predecessors in the realm of CV, facilitating advancements in a variety of application scenarios, such as advertisement replacement (Voulodimos et al., 2018; Chai et al., 2021). Convolutional Neural Networks, Deep Boltzmann Machines, Deep Belief Networks, and Stacked Denoising Autoencoders have been instrumental in addressing intricate CV problems. This has created avenues for innovative applications, including object detection, face recognition, action and activity recognition, and human pose estimation (Voulodimos et al., 2018). These advances also propel forward the field of ad replacement.

Despite the considerable progress in utilizing DL and CV in the domain of advertisement replacement, further enhancement remains possible. For instance, Ghassab et al. (2020) advanced a novel framework to automate advertisement replacement in soccer videos using deep learning strategies. Similarly, Yang et al. (2021) presented a method utilizing deep CNNs to replace Ads automatically in social media videos. Lee and Moon (2021) proposed an intelligent digital signage system based on edge computing and an intelligently trained advertisement recommendation model. Even with these advancements, there is still a need for extensive research to fully understand how to optimally deploy these technologies in real-world settings (Bara et al., 2022). Consequently, this thesis aims to delve deeper into the practical applications of DL and CV in advertisement replacement, thereby providing crucial insights into this evolving subfield. This solution will be accomplished through three stages: fence board

detection utilizing Key-point RCNN, Ads location tracking employing FlowNet and addressing human occlusion challenges using MODNet.

2.4 Overview of CNN

Artificial Neural Networks, or ANNs, represent a computational model inspired by the human brain's neural architecture. ANNs consist of a large number of interconnected nodes (neurons) or processing elements (PE) connected with coefficients (weights), analogous to the neural structure of the brain. Each connection between nodes can transmit a signal from one neuron to another. The receiving neuron processes the signal and signals downstream neurons connected to it (Agatonovic-Kustrin & Beresford, 2000). Over the years, they have proved instrumental in executing complex tasks in various fields, including medicine, finance, human resources, customer service, and supply chain (Jantan et al., 2018).

CNNs are a variant of artificial neural networks that have shown great efficiency and performance in various ML tasks. CNN has been particularly effective in image-driven pattern recognition tasks (O'Shea & Nash, 2015). The architecture of CNN typically comprises three main types of layers: Convolutional layer, Pooling layer, and Fully Connected (FC) layer. These layers help CNN to increase their complexity, identify greater portions of an image, and focus on simple features like colours and edges (Jinzhu et al., 2021).

An integral part of CV is object detection, and CNN has proven to be highly effective in this task (ZhiQiang & Jun 2017). CNN has been used in various applications such as pose estimation, vehicle detection, surveillance, and self-driving cars (Gupta et al., 2021). Region proposal base object detection algorithms, such as R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN, operate by trying to draw a bounding box around the object of interest within the image and are designed to adaptively learn spatial hierarchies of features through backpropagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers(Sood et al., 2021).

In the field of biological research, CNN has provided valuable insights. Scientists and clinicians have turned to these artificial intelligence systems to provide solutions for therapeutic development (Vaz & Balaji, 2021). Moreover, CNN is especially effective in the field of CV and object detection. It has been utilized for different applications such as image and video recognition, object detection, image classification, image segmentation, face recognition, signal identification, and time series prediction (Li et al., 2022).

R-CNN

The method of bounding-box object detection has been significantly advanced by the introduction of the Region-based Convolutional Neural Network (R-CNN) approach (Girshick et al., 2014). It utilizes a strategic focus on a feasible number of candidate object regions (Uijlings et al., 2013; Hosang et al., 2016), enabling an independent evaluation of each Region of Interest (RoI) by convolutional networks (LeCun et al., 1989; Krizhevsky et al., 2012).

Fast R-CNN

Fast R-CNN, proposed by (Ren et al., 2015), improved on the initial R-CNN by integrating region proposal and feature extraction into a single process. The method uses a Region of Interest (RoI) pooling layer to extract features from the proposed regions directly from the feature maps generated by the CNN. This allowed Fast R-CNN to train a deep VGG16 network nine times faster than R-CNN and increased test-time speed by 213(Girshick, 2015).

Faster R-CNN

Introduced by Ren et al. (2015), Faster R-CNN presents a pivotal development in object detection. This model incorporated a Region Proposal Network (RPN) to suggest regions for classification, substituting the selective search algorithm in its precursors. The result was a highly efficient, trainable process for proposing regions, lending itself to the moniker 'Faster' R-CNN. As a testament to its versatility and robustness, Faster R-CNN has been the foundation for several successful enhancements and currently leads in multiple benchmarks. For instance, Shrivastava et al. (2016) introduced a remarkably effective online hard example mining (OHEM) algorithm for training region-based ConvNet detectors based on this model. Furthermore, Lin et al. (2017) capitalized on the inherent multi-scale, pyramidal hierarchy of deep convolutional networks to build feature pyramids with minimal extra cost. Moreover, Huang et al. (2017) provided a unified implementation of Faster R-CNN, R-FCN (Dai et al., 2016), and SSD (W. Liu et al., 2016) systems. These they regard as meta-architectures and use to delineate the trade-off curve between speed and accuracy, achieved by varying key parameters such as feature extractors and image size within each meta-architecture.

Building on the Faster R-CNN architecture, He et al. introduced Mask R-CNN in 2017, extending object detection to instance segmentation. The method added a branch for predicting a binary mask for each RoI alongside the existing branch for classifying RoIs and predicting bounding boxes. This allowed Mask R-CNN to efficiently detect objects in an image while simultaneously generating high-quality segmentation masks for each instance (He et al., 2017).

2.5 Overview and Application of Key-Point Detection

The R-CNN family of algorithms has remained a significant area of interest within CV research over the past decade, specifically concerning object detection and keypoint identification. The Fast R-CNN model, introduced by Girshick (2015), drastically enhanced object detection performance relative to its predecessors. One notable improvement was the consolidation of the proposal and detection networks into a single network, leading to remarkable gains in accuracy and efficiency. Building on this success, Ren et al. (2017) proposed Faster R-CNN, incorporating a RPN into the model, thus facilitating near real-time object detection. The introduction and application of these models have expedited the evolution of object detection techniques. With the continued exploration of this field, the keypoint detection model emerged, offering a new perspective on object detection. This model can extract powerful 3D features from images and is hence deemed an essential resource for understanding 3D geometry. Parallel advancements were witnessed in related fields. For instance, Ma et al. (2017) proposed a multi-view 3D object detection network for autonomous driving, exemplifying another possible application of the keypoint detection model. Simultaneously, Zhou and Yu (2022) proposed variations of the R-CNN, such as the Point R-CNN, which offered an angle-free framework for rotated object detection, illustrating the adaptability of the base R-CNN framework for addressing specific challenges.

Faster R-CNN laid the groundwork for subsequent advancements in keypoint detection, giving rise to the Mask R-CNN and the Keypoint R-CNN. Notably, the Keypoint R-CNN is an extension of the Mask R-CNN, which builds upon the Faster R-CNN, exemplifying the iterative progress within the field. The Keypoint R-CNN was initially developed as an extension of instance segmentation for keypoint detection, primarily targeting object detection and human pose estimation (Ding et al., 2020; Linna et al., 2018). Given its ability to detect the corner point of objects, it became evident that the model could also effectively detect objects' backgrounds and, thus, be applied to advertisement replacement.

2.6 Overview of Optical Flow Measurement

The estimation of optical flow, which refers to the apparent motion pattern of objects between successive image frames, has made significant strides in recent years. Historically, many techniques, such as differential methods, region-based matching, energy-based techniques, and phase-based techniques, have been employed to compute optical flow (Horn & Schunck, 1981; Simoncelli et al., 2002; M. J. Black et al., 2002; Barron et al., 2003; Brox et al., 2004; Brox & Malik, 2011). Some of those techniques harness two critical constraints on image motion: data conservation and spatial coherence (Black & Anandan, 2002). Negahdaripour and Yu (2002) proposed an image motion constraint equation predicated on a model that permits the brightness of a scene point to vary over time, contributing to optical flow computation. Building upon these principles, Wedel et al. (2009) introduced an algorithm hinging on sequential convex optimization. Notably, this algorithm was capable of real-time processing and surpassed all previously published algorithms by over one average rank on the Middlebury optical flow benchmark, marking a significant progression in the field.

Despite their efficacy, modern deep learning techniques have progressively replaced traditional methods, specifically CNNs (Shah & Xuezhi, 2021). The task of optical flow estimation, which is essential in various image processing techniques, has witnessed significant advancements and a shift from traditional approaches to CNN-based methods (Shah & Xuezhi, 2021).

The introduction of FlowNet by Dosovitskiy et al. (2015) was a significant milestone in the field of optical flow estimation. FlowNet proposed using CNNs to directly learn the intricate mappings between image pixels and optical flow vectors, as opposed to previous methods that extensively relied on handcrafted features and energy minimisation techniques. This significantly departed from the typical association between CNNs and recognition tasks. FlowNet presented two novel architectures: a basic yet adaptable network structure and an advanced version with an improved performance correlation layer. The success of FlowNet can also be attributed to its innovative data training strategy. Dosovitskiy et al. (2015) developed a synthetic dataset titled "Flying Chairs," which proved effective for training the network to address the shortage of ground truth data for training. Notably, networks trained on this synthetic dataset demonstrated robust generalisation to real-world datasets, demonstrating the viability of training deep neural networks with synthetic data.

Dosovitskiy et al. (2015) spearheaded a revolutionary approach in the field, paving the way for the development and implementation of CNN-based methods for optical flow estimation. This groundbreaking work spurred investigations and evaluations of various CNN-based optical flow methodologies (Shah et al., 2020). For instance, Hui et al. (2018) introduced an

alternative network called Lite FlowNet. Despite being 30 times smaller in model size and 1.36 times faster in operation, Lite FlowNet matches the performance of FlowNet2 on the rigorous Sintel final pass and KITTI benchmarks. Similarly, Ranjan and Black (2017) proposed an innovative approach to compute optical flow, merging a classical spatial-pyramid formulation with deep learning. Moreover, Sun et al. (2018) developed a compact yet effective CNN model for optical flow, the PWC-Net. This model is 17 times smaller and more accessible to train than the recent FlowNet2 model.

Despite the growing prominence of CNN-based approaches, variational methods have also maintained a significant interest among researchers. For example, Tu et al. (2019) scrutinized the applications of variational and CNN-based methods in a systematic review of optical flow techniques, showcasing the broad range of methods currently being utilized in the field.

2.7 Overview of Human Segmentation and Matting

Traditional image matting techniques encountered difficulties like image distortion, uneven local details, and the requirement for additional input trimaps. Traditional image matting methods can generally be categorized into three types: sampling-based methods, propagation-based methods, and learning-based methods (L. Huang et al., 2023). Sampling-based methods predict the opacity of unknown regions by collecting a series of pixels or pixel blocks from the known areas of the trimap map. On the other hand, propagation-based methods predict the opacity of each pixel in unknown regions by establishing a connection between adjacent pixels and then using an optimization strategy to propagate the opacity information from known to unknown areas. Lastly, learning-based methods use large datasets to learn the image's features, which are then utilized to predict opacity (Wang & Cohen, 2007; Yao, 2017; Boda & Pandya, 2018; Li et al., 2019).

To address these issues, Xie et al. (2023) proposed an end-to-end network for bust matting, which improves edge refinement and global prediction for half-figure portraits. Focusing on uncertain regions at the bust's margins, the Contour Sharpness Refining Network (CSRN) was introduced to generate alpha maps with intricate edges. In spite of these enhancements, the technique still required manual annotation, prompting further research into semi-supervised solutions.

Portrait matting, which involves separating an image's object from its background, has recently received considerable attention (Ke et al., 2022; Xie et al., 2023; Zhang et al., 2022). The developments in this field have paved the way for applications in diverse fields, including photography, filmmaking, and AR (Ke et al., 2022).

The domain of portrait matting heavily relied on annotated images for learning, but detailed datasets and the manual annotation procedure were time-consuming (Zhang et al., 2022). Further advancements were achieved with the introduction of the Matting Objective Decomposition Network (MODNet) (Ke et al., 2022). MODNet was designed for portrait matting in real-time without the need for additional inputs or trimaps. Multiple sub-objectives were optimised concurrently using explicit constraints, and novel techniques such as Efficient Atrous Spatial Pyramid Pooling (e-ASPP) and self-supervised sub-objectives consistency (SOC) strategy were implemented. This method outperformed previous trimap-free methods, including the Adobe Matting Dataset and the Photographic Portrait Matting benchmark (PPM-100) (Ke et al., 2022).

In summary, advancements in portrait matting have evolved from traditional methods with manual annotations to semi-supervised techniques and, more recently, real-time trimap-free solutions.

Chapter 3: Methodology

The methodology for the advertisement replacement system in sports videos involves an intricate integration of three distinct but complementary deep learning models - Key-point RCNN-PF, FlowNet, and MODNet. The solution works in three stages: detecting fence boards using Key-point RCNN-PF, tracking the position of ads using FlowNet, and handling human occlusion using MODNet.

3.1 Fence board detection using Key-point RCNN-PF

3.1.1 Introduction to Key-point RCNN-PF Architecture

Key-point Region-Based Convolutional Neural Network (Key-point RCNN) is a state-of-the-art deep learning model designed for object and key point detection. It extends upon the Faster RCNN, a two-stage detector, by adding a parallel branch for predicting an object's key points alongside bounding box recognition. The Key-point RCNN - Picture Frame (Key-point RCNN-PF) (Zhang et al., 2020) constitutes an enhancement to the Key-point RCNN, with its specific function being the detection of vertices within picture frames. The primary motivation for choosing Key-point RCNN-PF for this project is its proven efficiency in detecting corner points of picture frames on objects of interest, which in this context, are the fence boards of sports games videos.

Key-point RCNN-PF has these main components: a backbone network, a Region Proposal Network (RPN), ROI Align and the Feature Pyramid Network (FPN). The backbone network is responsible for processing images and generating feature maps. The RPN generates object proposals and the ROI Align crops and aligns the Region of Interest (ROI) for key-point detection, while the FPN refines these proposals and makes final predictions. The reason to propose the Key-point RCNN is the addition of a Picture Frame branch (PF branch) of Key-point RCNN that predicts four vertices for each ROI, effectively enabling our algorithm to identify the locations of fence boards in the videos.

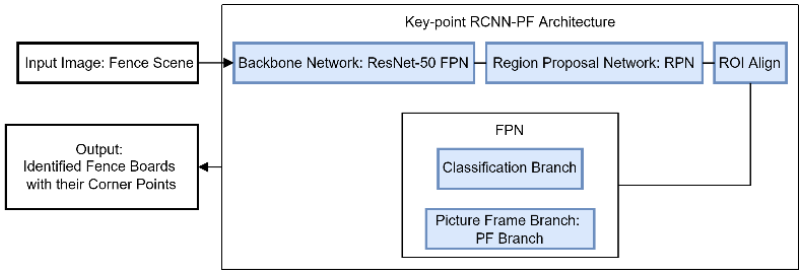

Figure 3.1 Workflow diagram of the Key-point RCNN-RF model

The Backbone Network as shown in Figure 3.1, in this case, is a ResNet-50 with FPN. ResNet-50 is a deep convolutional neural network with 50 layers and is known for its 'residual learning' technique that allows it to avoid the problem of vanishing gradients during training. It processes input images and generates feature maps, which are rich, high-dimensional representations of the input images' features. These feature maps are then used by the subsequent layers for further operations.

The Region Proposal Network (RPN) as shown in Figure 3.1, is a critical component of the RCNN architecture. In its operational process, the model examines the feature maps produced by the backbone network and proposals possible Regions of Interest (ROIs). These ROIs signify potential areas within the image where objects may be located, and they are subsequently utilized in the subsequent ROI Align stage. These proposals are based on different scales and aspect ratios to accommodate objects of varying sizes and shapes. Each proposed region is given an 'objectness'' score, which estimates the likelihood of the region containing an object.

Once the RPN formulates the region proposals, the subsequent operation of ROI Align as shown in Figure 3.1, is initiated. It extracts the proposed regions and performs a kind of 'fine-tuning' to align them more precisely. This involves cropping and resizing the regions to a fixed size to be processed by a fully connected layer. Unlike older methods like ROI pooling, ROI Align uses bilinear interpolation to maintain spatial precision, thereby preventing the warping of features during the cropping and resizing process. This improves the model's performance, especially when handling small objects. The alignment process corrects the misalignments caused by the quantization process in RPN and thus leads to more accurate predictions.

3.1.2 Feature Pyramid Network (FPN)

The FPN as shown in Figure 3.1, is a structure that uses a top-down architecture with lateral connections to build an in-network feature pyramid. The purpose of this architecture is to efficiently construct a prosperous, multi-scale feature representation that can cater to objects of

different sizes. In the Key-point RCNN-PF, FPN takes the feature maps from the ResNet-50 backbone, applies multiple convolutional layers to create a pyramid of features at multiple scales, and then refines the region proposals made by the RPN for making the final predictions.

The Picture Frame branch (PF branch) (Zheng et al., 2021) is an additional component introduced in the Key-point RCNN-PF architecture. It is designed to detect better the vertices of picture frames, which are small-scale features usually located at the edges of objects. The PF branch adopts strategies such as increasing the number of channels in the layers, using convolutional layers for down sampling, and incorporating an FPN structure for handling multi-scale detections. After the ROI Align operation, the PF branch uses convolution and deconvolution layers, lateral connections, and a key-point predictor to produce the final output. This results in more accurate detection of picture frames and their vertices.

3.1.3 Dataset Preparation, Model Training, and Fine-Tuning for Fence Board Detection

Preparing a robust dataset is essential in training a deep learning model. For the task of fence board detection, a full tennis game competition video was chosen to serve as the primary source for our dataset. The game was diverse in its shots, angles, and lighting conditions, making it an ideal choice for gathering varied data. Some of the clips were chosen as the introductory demo video for this project. The FFmpeg software was utilized to extract diverse frames from various game periods, offering a wide variety of visuals for the model. This included both close-ups of players, wide-angle views of the court under varying lighting conditions, and different-angle views of the court. These frames were then selected to create the training and testing datasets. LabelMe, an open-source graphical image annotation tool, was used to annotate the images in the training dataset. The process involved manually marking key-points on the fence boards within the images, which helped in teaching the model about the boards' shape and location.

The Key-point RCNN-PF model was trained using a carefully curated dataset in an iterative process that optimized its internal parameters based on prediction error. This involved multiple runs or epochs, and the learning rate was periodically adjusted to enhance the training process. A specific learning rate was identified to be the most effective in allowing the model to learn at a steady pace without overshooting the optimal solution. Fine-tuning the model weights was also an essential part of the process. A point in the training process was identified where the model's weights provided the highest accuracy in detecting the target objects. Additional images representing exceptional viewing angles were included in the training dataset to ensure more comprehensive learning and to improve the robustness of the model.

The specifics surrounding the dataset assembly and annotation, along with the process of training and fine-tuning the Key-point RCNN model, will be elaborated upon in the forthcoming Experiment section.

The model was continually fine-tuned throughout training by adjusting parameters and adding new data to ensure optimal performance. This rigorous training approach helped us mitigate the challenge of automated fence board detection in sports video frames.

3.2 Ads location tracking using FlowNet

The task of tracking the location of Ads in sports games videos necessitates the use of a robust and efficient model capable of understanding and predicting the motion of objects within video sequences. One such model is FlowNet, an advanced deep-learning architecture specifically designed for optical flow estimation. FlowNet's key strength lies in its ability to learn to recognise and predict intricate motion patterns from large-scale data sets, making it exceptionally competent in tasks like tracking the location of Ads in videos. Optical flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and the scene. The use of optical flow is fundamental to this task, primarily due to the unique positioning of the fence board in the video frames. As the fence board is located in the video's background, its movement patterns mirror the general background motion. These subtle movements can be effectively captured by employing optical flow measurement, enabling the model to track and analyse the changes in position and orientation of the fence board over time. FlowNet provides a robust and computationally efficient solution for tracking the position of specific objects, such as fence boards, across a sequence of frames in a video, and that is the reason to propose that Flownet track the Ads location.

3.2.1 Introduction to Optical Flow

Optical flow is a task of immense importance in the field of CV and image processing. It involves estimating the apparent motion of every pixel in an image between consecutive video frames. Doing so allows for understanding and predicting the movement of objects, surfaces, and edges in the visual scene based on variations in intensity patterns over time. The applications of optical flow are numerous and significant, ranging from object detection and tracking, video compression, and activity recognition to video analysis and more. These applications are critical in the context of the project, necessitating a deep understanding of

optical flow. The optical flow estimation significantly improves the accuracy of fence board detection and alignment across consecutive frames.

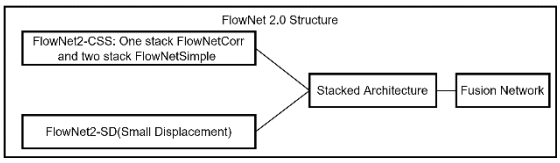

Figure 3.2 The architecture of FlowNet 2.0

3.2.2 Overview of the Convolutional Neural Network (CNN) Architecture-FlowNet

FlowNet represents a pivotal shift in CV tasks, specifically in the realm of optical flow estimation, by harnessing the power of Convolutional Neural Networks (CNNs). It represents a break from past instances where CNNs had not been efficiently applied in the field of optical flow estimation (Dosovitskiy et al., 2015). The architecture of FlowNet implements a symmetrical encoder-decoder structure. The encoder part progressively reduces the spatial dimension and increases the feature dimension, resulting in a compact feature representation. This is followed by the decoder part, which gradually recovers the spatial resolution to eventually generate a dense optical flow field corresponding to the input image pair.

The design of FlowNet involves the deployment of CNNs in a way that treats optical flow estimation as a supervised learning problem. Two architectures (Dosovitskiy et al., 2015) were proposed in its conception. The first, named FlowNetSimple, is a generic architecture that only consists of convolutional layers, and the second, called FlowNetCorr, integrates a unique layer named the correlation layer, which is used to perform multiplicative patch comparisons between two feature maps that creates correlations between feature vectors across various image locations.

FlowNetCorr is a variant of the original FlowNet model that introduces a correlation layer to the architecture to calculate matching scores between two feature maps. The architecture consists of two CNNs, one for the target frame and the other for the reference frame. The correlation layer proves pivotal in the computation of optical flow as it helps establish correspondences between the two frames, thus enabling more accurate and robust motion tracking.

The FlowNetSimple model simplifies the FlowNetCorr architecture by substituting the correlation layer with a more straightforward method of simply concatenating feature maps from two consecutive frames. This modification substantially reduces the computational

complexity of the model while ensuring that the effectiveness of optical flow estimation is not compromised.

FlowNet 2.0 as shown in Figure 3.2, represents a more advanced end-to-end architecture that incorporates multiple FlowNetSimple and FlowNetCorr layers stacked together in a cascading manner. In FlowNetSimple, two sequentially adjacent images are stacked together as input, while FlowNetCorr separately convolutes two images and then fuses them using a correlation layer (Ilg et al., 2017). This model combines the strengths of FlowNetSimple and FlowNetCorr, yielding better performance than either model when used independently. Specifically, FlowNet2.0 harnesses the small displacements learned by FlowNetSimple and the large displacements known by FlowNetCorr to generate a more accurate optical flow field.

3.3 Human Occlusion Solution using MODNet

3.3.1 Overview of MODNet

Ke et al. (2022)'s Matting Object Decomposition Network (MODNet) revolutionises the field of digital image matting by specialising in real-time, trimap-free portrait matting. MODNet assists in separating the foreground from the background in an image, a process called image matting. Conventionally, trimaps are used to achieve this separation, but the process is laborious and resource intensive. MODNet, however, provides an alternative devoid of trimap, thereby increasing the accessibility and potential applications of the image matting process.

Architecturally, MODNet comprises a low-resolution semantic estimation branch and a high-resolution detail prediction branch, in which the outputs of these branches are integrated to yield the final alpha matte. This design enables MODNet to handle tasks involving semantic estimation and detail prediction concurrently and accurately.

3.3.2 Architecture of MODNet

MODNet circumvents this trimap pre-requisite by implementing an end-to-end network architecture for trimap-free portrait matting. The MODNet achieves trimap-free matting due to its unique architecture, encompassing three key components: semantic estimation, detail prediction, and semantic-detail fusion. Initially, the semantic estimation module generates a broad understanding of the image's structure, thus eliminating the need for a predefined trimap. Subsequently, the detail prediction module identifies intricate details in the image, providing

precise edge information and reducing reliance on trimaps. Ultimately, the semantic-detail fusion module merges the global and local insights of the image to produce a refined, high-resolution alpha matte. The blend of broad and focused perspectives in MODNet's architecture facilitates accurate object segmentation without trimaps. This trimap-free approach makes it adaptable for real-time interactive applications where the creation and manipulation of trimaps are impractical.

Semantic estimation in MODNet is performed by the low-resolution branch S. Instead of employing a separate decoder with a large parameter set, MODNet extracts high-level semantics via an encoder. This approach makes the semantic estimation process more efficient, efficiently locates the portrait in the image, removing the need for a separate decoder and provides valuable high-level representation useful for subsequent branches and joint optimization. The performance of semantic mask prediction is enhanced using an efficient variant of Atrous Spatial Pyramid Pooling (ASPP) - known as e-ASPP. The conventional ASPP leverages multiple parallel atrous convolutions with different rates to capture multi-scale contextual information. However, its computational overhead is significant. The e-ASPP module introduced in MODNet overcomes this issue by adjusting the original ASPP design, which enhances the handling of multi-scale feature extraction and fusion, substantially reducing computational demand without sacrificing performance.

Detail prediction in MODNet is made through the high-resolution branch D, which processes a transition region surrounding the foreground portrait by reusing the low-level features to reduce computational overhead. This detail prediction branch leverages the high-level portrait semantics as a priori for detail prediction, resulting in a detailed boundary matte.

Semantic-detail fusion is the final step in MODNet, where the outputs from the semantic and detail branches are combined to predict the final alpha matte. This fusion process measures the absolute difference between the input and the composited images obtained from the predicted alpha matte α_p and the ground truth (foreground and background) α_g, which is optimized by the customized loss function $\mathcal{L}_\alpha$, controlled by a compositional loss $\mathcal{L}_c$, as Eq. (3-1):

$$\mathcal{L}_\alpha = \| \alpha_p - \alpha_g \|_1 + \mathcal{L}_c \tag{3-1}$$

SOC is a strategic component of MODNet designed to maintain the consistency of output among the sub-objectives of the model. This strategy imposes consistency constraints among predictions, ensuring a coherent representation of portrait semantics and details across different model components. By leveraging this self-supervised SOC strategy, MODNet can adapt to unseen data distributions, thus enhancing the model's performance and ability to adapt to unseen data distributions.

3.3.3 Adapting MODNet for Occlusion Handling in Sports Videos

The MODNet model, known for its precision in trimap-free image matting, was adapted to handle human occlusions in sports videos. The adaptation process entailed meticulous data collection and preparation, rigorous training of the model, and continuous evaluation and fine-tuning to ensure optimal performance.

Figure 3.3 Example of MODNet Training Data

In the process of addressing human occlusion issues, meticulous data collection and preparation were undertaken. This involved sourcing a multitude of images where tennis players and ball picker were positioned in front of the fence boards, with additional images showcasing players near various display boards, such as the green time display board and blue speed display board, as shown in Figure 3.3. Each image was carefully annotated using LabelMe to mark the boundaries of humans and key elements in the scene. These annotations were critical in aiding the model to comprehend the shapes and locations of human subjects. The training images were further processed to create binary masks and alpha mattes. Each image was given a binary mask where the human subject was represented in red and the background in black, as shown in Figure 3.4. Simultaneously, an alpha matte was created for each image to represent the transparency level of each pixel， in which the background is 0 and the human is 255, where the human subject was represented in white and the background in black, as shown in Figure 3.4. The dataset was further augmented by generating binary masks and alpha mattes for each image, assisting MODNet in accurately segmenting human figures.

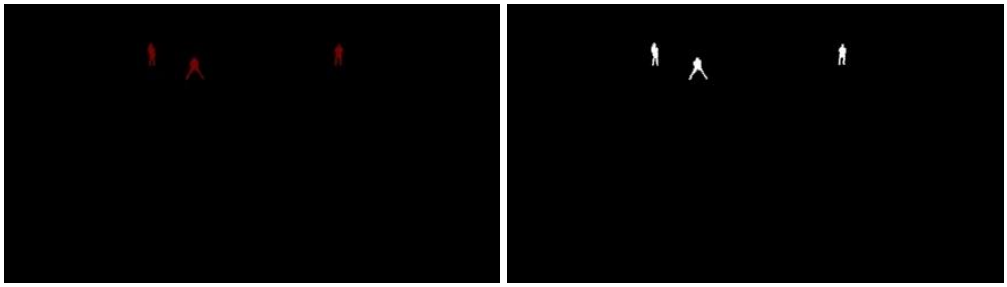

Figure 3.4 Example of Binary Mask (Left) and Alpha Matte (Right)

In the procedure of image preprocessing, the generation of alpha mattes from preexisting masks is a critical task. The transformation involves a systematic sequence of steps outlined as follows: Initially, the transformation operates by sequentially accessing and loading each mask file from a predetermined directory. Once an image is retrieved, it is subsequently converted into a binary format, which means every pixel is marked as either "on" or "off.". This binary conversion clarifies each pixel as active or inactive, simplifying the image data. Post binary conversion, the algorithm meticulously extracts the single channel from the binary image, thus allowing a focus on portions of the image that predominantly possess blue colour information. After this extraction process, the edge detection phase is initiated. For this purpose, the Canny edge detection technique is employed, which can distinguish the edges or boundaries between distinct objects and the background within the image. Upon successfully detecting edges, the algorithm proceeds to map the contours, or the perimeters, of these detected boundaries. Once identified, these contours are graphically represented on the original image, yielding an alpha matte, where the value is 128. This matte accentuates human figures, along with their encompassing shapes.

In essence, alpha mattes furnish information regarding the level of transparency for each pixel present in a given image. Applying alpha matte is highly pertinent to human segmentation as it is instrumental in restricting human subjects from the encompassing background. During the pivotal training phase, the MODNet is designed to predict the alpha matte for a designated image. The alpha matte assigns a continuum of values from 0 to 1, with 0 signifying the background (absolute transparency) and 1 indicating the foreground (absolute opacity), and values such as 0.5 in between representing the level of transparency. This transparency gradient is particularly helpful in dealing with difficult areas such as hair or semi-transparent objects where foreground and background are not binary. Unlike binary masks that provide a hard partition, alpha mattes provision of a soft partition between the foreground and the background. This facilitates the capture of nuances in the boundary definitions of human subjects, which may be blurred due to factors such as motion or lighting conditions. Hence, alpha mattes demonstrate their indispensability by providing a nuanced approach to the depiction of object transparency within an image. This assists the MODNet model in producing

highly precise human segmentation, enhancing the output's accuracy and detail. This unique ability to transcend the rigidity of binary segmentations and capture real-world imagery's finer details underscores alpha mattes' importance when doing human segmentation with the MODNet model.

An iterative procedure was adopted in the subsequent phase of training and fine-tuning MODNet. The model was subjected to numerous training epochs, adjusting its internal parameters based on the errors in its predictions. Key hyperparameters, such as learning rate and batch size, were carefully selected to facilitate stable learning and prevent overstepping the optimal solution. After repeated cycles of training and evaluation, the model's weight at a specific epoch demonstrated superior performance in human segmentation. This thorough process of training, evaluation, and fine-tuning ensured that MODNet was proficient in managing human occlusions in sports videos, enhancing the effectiveness of the Ads replacement system.

3.4 Integration of the Three Stages

Our methodology revolves around the ingenious fusion of MODNet, Key-point RCNN, and FlowNet, three state-of-the-art deep learning models, each addressing a particular challenge in the advertisement replacement task within sports videos. Specifically, Key-point RCNN-PF adeptly detects fence boards, FlowNet skilfully tracks the Ads locations in the video frames, and MODNet expertly addresses human occlusions. The integration of these models allows them to capitalize on their strengths while creating a unified system that is more than the sum of its parts. These three models work harmoniously, creating a cohesive and efficient advertisement replacement system that demonstrates unprecedented capabilities in handling the complexities of sports videos.

3.4.1 Preparations for the three models

3.4.1.1 Fence board Detection and Ads pasting

The initial stage of this complex operation is delegated to the Key-point RCNN-PF model, a complex architecture whose primary function is to detect the presence of fence panels in sports video frames. In performing this critical task, it provides a comprehensive understanding of the structure of the fence panels, pinpointing their precise location and accurately delineating their boundaries.

Figure 3.5 Example of Annotated Training Image

The Figure 3.5 presents a representative instance of the training data annotated using the Labelme tool. The boundary of the fence board is demarcated by a rectangular frame in red, which also pinpoints the four corner key points integral to its geometry. Simultaneously, a green rectangular frame highlights the region within which the fence board is detected, providing a visual reference for its spatial positioning.

The Key-point RCNN-PF model achieves this by carefully deploying numerous interconnected layers to scan video frames. These layers are able to recognize the tiniest changes in pixel values, allowing them to detect the different shapes, patterns and colours of fence board features. In addition to pure detection, the model further identifies key points on the detected fence board to establish their geometric and spatial context within the frame. This critical step not only considered the fence panels' physical characteristics but also placed them within the larger visual structure of the frame, incorporating aspects such as perspective, orientation and relative positioning.

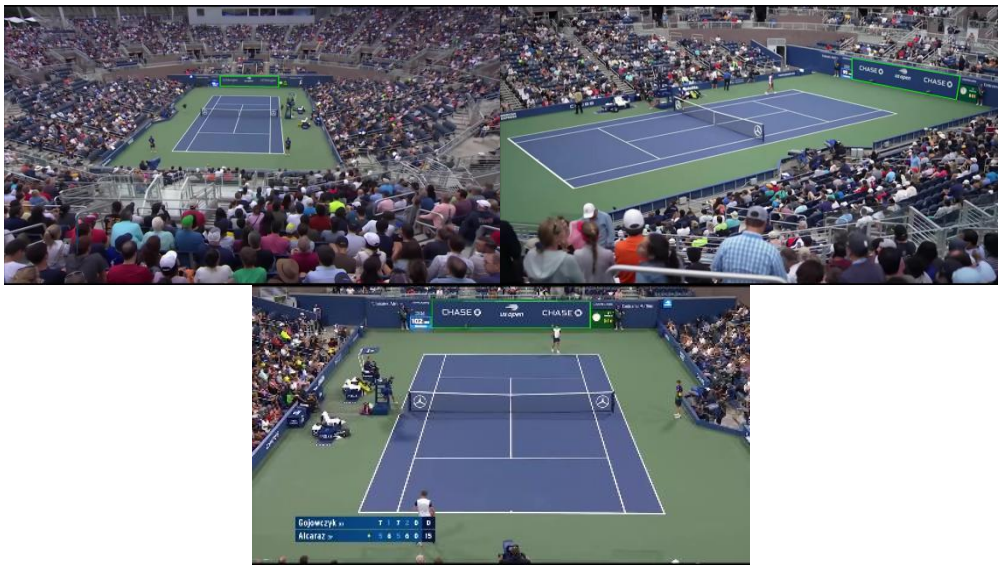

Figure 3.6 Example Results of the Key-point RCNN-PF Model

The above Figure 3.6 indicate the resultant output generated by the Key-point RCNN-PF model. Judging from the depicted outcomes, the model's training appears to have met the

anticipated goals, successfully discerning the area of the fence board along with the crucial four corner key points under diverse perspectives. This capability extends to various situations encompassing long-distance captures, proximate close-ups, and shots taken at varying angles, demonstrating the model's versatility and robustness in object detection.

The fundamental basis for the subsequent process is the initial detection and localization of the fence panels completed by the Key-point RCNN-PF model. It essentially maps the environment in which these processes operate, providing them with the necessary data and context to perform their respective functions accurately and efficiently. Therefore, the role of the keypoint RCNN-PF model in this process goes far beyond simple detection, directly affecting and enhancing the later performance of the comprehensive system. Upon attaining the anticipated outcomes from the model's training, a solid foundation has been established to proceed with the Ads pasting function.

The incorporation of Ads into the video frames is achieved through a systematic process that leverages image manipulation techniques and hardware-specific optimizations. The primary stage of this process is loading the advertisement image from a predetermined path. This phase is crucial to converting the image's colour space from the Blue, Green, Red (BGR) format to the more universally employed Red, Green, Blue (RGB) format, ensuring the image is compatible with standard image processing routines. After image retrieval, the function pasting Ads is invoked to superimpose the advertisement image onto the video frame. The advertisement image is loaded, and integral properties such as its height, width, and the number of colour channels it comprises are identified. Concurrently, the corner coordinates of the advertisement image and the target region on the video frame, destined for advertisement placement, are precisely determined. These corner points empower the calculation of a transformation matrix, computed using the corner points with the cv. getPerspectiveTransform function, which effectively manipulates the perspective of the advertisement image to align with that of the target region on the video frame. This transformation is subsequently applied to the advertisement image. Finally, Replace the appropriate pixels in the video frame with those corresponding to the ad image. The culmination of this intricate procedure is a video frame that flawlessly incorporates the advertisement in the predetermined area. This method's efficacy largely depends on the precise transformation of the advertisement image, enabling it to be seamlessly integrated within the target area on the video frame.

3.4.1.2 Human Occlusions Solution

Figure 3.7 Example of Human Occlusion (Left) and Solution (Right)

As shown on the left side of the above Figure 3.7, the tennis player in front of the fence board is blocked when the advertisement is pasted, so a solution to this occlusion problem is needed. The right side of the Figure 3.7 above is the result after applying the solution.

Upon successful identification and localization of the fence boards within the video frames through the Key-point RCNN-PF model, the MODNet model is activated. Its primary function in this context is to apply its comprehensively trained capabilities to separate foreground objects, specifically human figures, from the background in the video frames.

Figure 3.8 Example of the Comparison of the Original image (Left) and the Processed Image (Right)

The information encoded in the binary mask and the alpha matte, developed during the data preparation stage, MODNet accurately discriminates between the human figures and the background elements, such as the fence boards or nearby time and speed display boards. As a result, human figures, which may include athletes or ball pickers located in front of the boards, are effectively segmented. This ensures their visual prominence is maintained, irrespective of the subsequent insertion of Ads into the background, thus preserving the integrity of the original video footage. The Figure 3.8 displays a side-by-side comparison of the original image (on the left) and the processed image following the application of the MODNet model (on the right). The segmentation result is visually represented by rendering the human figures in a distinctive pink colour. Through this visual comparison, the effectiveness of the MODNet model in delineating the human figures from their surrounding environment, particularly in proximity to the fence board area, becomes readily apparent.

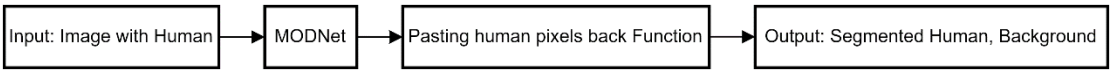

Figure 3.9 Workflow Diagram of Human Occlusions Solution

This cutting-edge implementation of the MODNet model offers a unique and efficient solution for real-time human occlusion handling in sports videos. It enhances the spectator's viewing experience by ensuring clear visibility of the athletes and provides an adaptable framework for integrating Ads seamlessly into the video. A vital component of this solution as shown in Figure 3.7, resides in the function designed to 'paste human pixels back', as shown in Figure 3.9. This operation distinguishes the human figure in front of the fence board from the background scenery. Once the advertisement has been integrated, this function replaces the segmented human pixels in their initial positions within the image. Consequently, even after advertisement implantation onto the fence board, the human images remain distinctly visible, maintaining the overall coherence of the original video content. This method ensures an unobtrusive advertising experience, where the Ads meld seamlessly into the environment. At the same time, the primary content - human figures in action - continues to engage viewers' attention without any visual disruption.

Additionally, once the issue of Human Occlusions is successfully mitigated and the devised solution exhibits the intended outcome, the flexibility of this model broadens its prospective applications, reaching beyond the confines of image matting. Specifically, within the scope of this dissertation, the focus is its amalgamation with fence board replacement algorithms and advertisement location tracking systems. The implications and challenges will be explored in more depth in the subsequent sections of this study.

3.4.1.3 Ads Location Tracking

Upon successfully solving the fence board detection, Ads pasting, and human occlusion simultaneously, FlowNet steps in to maintain the consistency of the advertisement placement throughout the video frames. By utilizing its strength in optical flow estimation, FlowNet effectively tracks the motion and continuity of the placed Ads across the video. This ensures that the Ads remain stable and well-integrated with the scene despite dynamic shifts in the scene or changes in camera perspectives. Without the application of FlowNet, the fence board would need to be detected frequently. Due to variations in each frame, the detected area of the fence board would be different, leading to a visually inconsistent rectangular pattern.

In the context of tracking the advertisement's location and ensuring its stability across video frames, the role of the optical flow field becomes particularly significant. The variable, the optical flow field, essentially constitutes a flow map that depicts the movement of each

pixel between consecutive frames in the video. It is organized as a 3-dimensional array where the first dimension represents the height of the image, the second dimension represents the width, and the third dimension represents the channels. In the case of optical flow, there are two channels, one for horizontal movement U and one for vertical movement V. As such, it plays a pivotal role in tracking and maintaining the consistency of the advertisement's position over time.

Given the input images denoted as $I_n(x, y)$ and $I_{n-1}(x, y)$, the FlowNet model is invoked to generate a corresponding optical flow map, represented as $P(x, y)$. This map is subsequently partitioned into two distinct constituents - the horizontal flow symbolized as $U(x, y)$, and the vertical flow, represented as $V(x, y)$. This strategic bifurcation allows for the nuanced interpretation of pixel movements along the horizontal and vertical axes, thereby providing a comprehensive understanding of the spatial dynamics encapsulated within the input image.

$$u_{n-1}^{i} = U\left(x_{n-1}^{i}, y_{n-1}^{i}\right) \tag{3-2}$$

$$v_{n-1}^{i} = V\left(x_{n-1}^{i}, y_{n-1}^{i}\right) \tag{3-3}$$

This signifies that the horizontal and vertical optical flow at the location of each keypoint in the previous frame (n-1), which are shown as Eq. (3-2) and Eq. (3-3), is utilized to determine the predicted speed for the keypoint in the current frame (n). The U and V represent the functions that provide the horizontal and vertical optical flows, respectively. The x_{n-1}^{i} and y_{n-1}^{i} represent the x and y coordinates of the keypoints at the previous $(n-1)$ frame. Here, n refers to the current frame, and i refers to the index of the keypoints.

In this procedure, the keypoints of the advertisement are initially identified and placed on the fence boards. These keypoints serve as pivotal references that mark the advertisement's location and orientation in each frame. The U and V vectors, derived from the $P(x, y)$, describe the x (horizontal) and y (vertical) displacements for each pixel and, crucially, for the keypoints representing the advertisement. By adding the x and y displacements to the previous keypoint coordinates, the predicted positions for the keypoints in the current are computed. These new positions represent where the advertisement should be in the current frame. The process is represented as Eq. (3-4) and Eq. (3-5):

$$x_{n}^{i} = x_{n-1}^{i} + u_{n-1}^{i} \tag{3-4}$$

$$y_{n}^{i} = y_{n-1}^{i} + v_{n-1}^{i} \tag{3-5}$$

where x_{n}^{i} and y_{n}^{i} represent the x and y coordinates of the keypoints at the current frame (n).

This motion-based prediction helps maintain the continuity and stability of the advertisement's position throughout the video. It ensures that the advertisement moves coherently with the rest of the scene, maintaining a stable position relative to the fence boards despite changes in perspective, camera movement, or dynamic shifts in the scene. This consistency creates a more seamless and natural advertisement integration into the video, enhancing the visual coherence and overall viewing experience.

3.4.2 Initial Integration Model

The final challenge our methodology addresses is the seamless integration of these three distinct models into a unified pipeline. This involves consolidating the Python scripts for each model into one system, requiring rigorous testing and fine-tuning of parameters to ensure the optimal functioning of each model without interference. Careful debugging and iterative adjustments have led to a robust Ads replacement system capable of effectively managing the complexities of sports videos and delivering visually appealing, high-quality results.

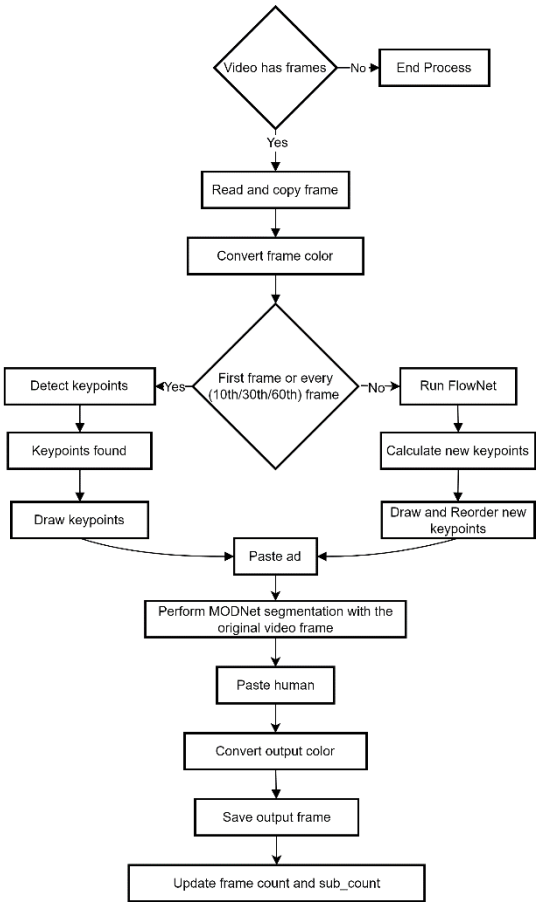

Figure 3.10 Workflow Diagram of Initial Integration Model

The Figure 3.10 is the initial integration model, which provides a proposed solution to video processing that includes advertisement insertion and maintains the visual continuity of

the scene. This procedure harnesses advanced CV techniques along with pre-trained deep learning models like keypoints detection (Key-point RCNN-PF), optical flow estimation (FlowNet) and human segmentation (MODNet).

In the preliminary processing phase, the model initially verifies the presence of frames within the input video. This is an essential prerequisite to ensure the model has visual data to process and act upon. In the positive event of frame existence, the model progresses to a crucial operation: transforming the colour space of the image frames from the Blue-Green-Red (BGR) to the Red-Green-Blue (RGB) format. The rationale behind this colour space conversion is primarily dictated by the unique operational framework of the OpenCV library, a fundamental component in our model's architecture. Diverging from the traditional conventions of image processing that predominantly use the RGB format, OpenCV uniquely operates within the BGR colour space. Consequently, to leverage the full suite of functionalities offered by OpenCV, the initial RGB images are transmuted into the BGR format, aligning the image representation with the library's operational standards and effectively setting the stage for subsequent stages of the processing workflow.

Following successfully converting the image colour, the model transitions to the frame verification stage. Here, a strategically instituted counter is used to ascertain whether the frame under consideration is the first in the input video or every 60th frame of the video. The choice of these particular frames (i.e., every 1^{st}, 60^{th}, 120^{th} ,180^{th}, etc.) is not arbitrary but is grounded in optimization strategies to enhance the model's overall performance. If a frame satisfies these conditions, the model Key-point RCNN-PF is activated. This model has been designed to detect and identify the keypoints associated with the fence board within the video frame. Upon successfully detecting the keypoints, a bounding box encapsulating these detected points is created and superimposed on the frame. However, in the contrasting scenario, where a frame does not satisfy the abovementioned conditions, such as when the fence board is not detected in the second to ninth frames, the model adopts a different approach, invoking the FlowNet model. The FlowNet model is provided with the current and the preceding frame as inputs to generate this optical flow map, referred to as P. Each pixel in this optical flow map carries a two-dimensional vector (U and V), signifying the pixel's estimated horizontal and vertical displacement from the previous frame to the current one.

After this, the model concentrates on updating the locations of the keypoints to reflect their positions in the current frame, positions that would dictate where the advertisement should be placed. Initially, the floating-point keypoint coordinates are converted into integer values to align with the image's pixel grid. The integer-format coordinates allow the extraction of their corresponding motion vectors U and V from the previously derived P. The model then applies

these vectors to predict the updated positions of these keypoints within the current frame, thereby ensuring the accurate placement and continuity of the advertisement across frames. Once the model has updated the keypoints' positions, it executes a crucial operation - the reordering of keypoints. Because the OpenCV function requires a different keypoint order than the keypoint detection, this step is more than just an organisational procedure; it’s fundamental to maintaining a uniform frame of reference for the placement of the advertisement within each frame. This reordering process adheres to a predefined sequence, ensuring a standardised approach across all frames, guaranteeing consistency in the advertisement's relative positioning.

After this reordering, the Ads pasting function is called into action. The role of this function is to facilitate the precise integration of the advertisement image within the current frame. Operating on the inputs of the reordered keypoints and the advertisement image, Ads pasting function undertakes a sophisticated image fusion process. The function subsequently outputs a frame which smoothly incorporates the advertisement, presenting a seamless visual amalgamation that maintains the integrity of the original video while effectively displaying the advertisement.

In an attempt to mitigate any visual disturbances potentially arising from overlaps between human figures in the frame and the newly inserted advertisement, the MODNet model is strategically deployed. MODNet, a proficient model in human image segmentation, facilitates the extraction of human figures from the original video frame, consequently generating a binary mask. This mask undergoes a critical thresholding process, in which the optimal threshold from the experiment is 128, effectively separating pixels associated with human figures from those pertaining to the background. Empowered by this binary mask, the paste human pixels back function is invoked. This function is critical in ensuring a visually coherent advertisement integration by overlaying the human figures onto the inserted advertisement. Working with advertisement images and the binary mask as inputs, the paste human pixels back function undertakes a complex blending operation, which culminates in generating a final output image where the advertisement is seamlessly integrated, and human figures are maintained in their foreground position.

Following the complex overlay operation, the output image is subjected to subsequent colour conversion, transitioning from the RGB format to the BGR format, as required by the OpenCV save video function. This transformation is pivotal in preparing the output image for display purposes and incorporation into the resulting output video file. The successful completion of these procedures signals the termination of a single iteration within the overall processing loop, with the current frame duly updated to serve as the 'previous frame' in the forthcoming iteration.

This initial integration model embodies an intricate strategy for accurate keypoint tracking within a video sequence, leveraging these tracked keypoints for the dynamic insertion of Ads into the scene. Additionally, a vital priority of the model is to guarantee that human figures within the frame uphold their dominant foreground status, notwithstanding the introduction of the advertisement. The ultimate composite image generated through this process is visually harmonious and exhibits seamless integration, presenting an advertisement that is flawlessly interwoven with the original scene, enhancing viewer engagement without disrupting the visual continuity of the footage.

Chapter 4: Experimental Results and Analysis

This chapter introduces the results and analysis of each deep learning model, along with their actual applications in the advertisement replacement system. The models examined include the Key-point Region-Based Convolutional Neural Network - Picture Frame (Key-point RCNN-PF), which is used for Fence Board Detection, the Matting Object Decomposition Network (MODNet), utilized for Human Occlusion Solution, and FlowNet, employed for Ads Location Tracking. Furthermore, the results and analysis of the integrated models' applications are displayed, encompassing the Integration of MODNet and Key-point RCNN-PF, the Initial Integration Model, and the Final Integration Model.

Here is the summary of the training and testing datasets used for each model; detailed information about each dataset is presented in the respective sections. The Key-point Region-Based Convolutional Neural Network - Picture Frame (Key-point RCNN-PF) model was trained with a dataset comprising 105 images and tested with a dataset of 26 images. Similarly, the Matting Object Decomposition Network (MODNet) model was trained with 150 images and tested with a set of 15 images.

4.1 Fence Board Detection

4.1.1 Dataset Preparation and Annotation

The dynamic nature of tennis matches and the myriad situations that can present themselves near the fence board were thoughtfully considered during the dataset selection process for model training. It was understood that scenes involving the fence board are not static, as they are subject to continuous change throughout the match, presenting many diverse conditions. These can include various camera angles that capture the fence board, moments when the fence board is not obstructed by foreground elements or instances where human activity occludes the fence board.

Figure 4.1 Example Training Data of the Key-point RCNN-PF Model

In response to these diverse circumstances, the training data was assembled to include the frames showcasing various situations near the fence board. This ensured that the model was exposed to and trained on a representative variety of conditions, thus enhancing its adaptability and robustness in managing the inherent variability of sports videos. A range of camera angles was incorporated into the training data, with frames capturing various perspectives such as close-ups, long-range shots, direct frontal views, and oblique side views, as shown in Figure 4.1. This diversity ensured that the model was trained to recognize the fence board and determine the correct placement of Ads regardless of the camera angle or the shot's distance.

Furthermore, frames depicting various scenarios near the fence board as shown in Figure 4.2, were also included. These ranged from scenarios where a tennis player stands in front of the fence board, creating potential occlusion, to situations where multiple individuals, like a tennis player and a ball picker, appear near the fence board simultaneously. Instances where a person only partially occludes the fence board by appearing on its edge were also included.

Figure 4.2 Example of Various Scenarios Near the Fence Board

To facilitate this diverse data collection, the software FFmpeg was employed to extract 969 frames from various periods of the game. This ensured a diverse dataset, with frames ranging from close-up shots of players to wide-angle court views and from daylight scenes to under-the-floodlight moments. From these 969 frames, 105 images were carefully selected for the training and 26 for the testing datasets. This selection was based on factors such as the clarity, angles, and visibility of the fence boards, which are crucial for the optimal functioning of the Key-point RCNN-PF model.

Providing these diverse scenarios in the training data allowed the model to develop a nuanced understanding of the complex dynamics of a tennis match. The model was trained to accurately detect the fence board irrespective of the situations unfolding near the fence board.

This approach to data selection and training ultimately improved the model's overall performance, enhancing its ability to integrate Ads into diverse sports video frames seamlessly.

4.1.2 Dataset Annotation

The images selected for the training dataset were annotated using LabelMe, an open-source graphical image annotation tool. The images were annotated and involved marking the key points on the fence boards manually. Each key point was selected to provide the model with an understanding of the shape and location of the fence boards. The below Figure 4.3 show how the fence board is labelled in different situations. The fence board will show different rectangle-like shapes at different angles and in different situations, so be very careful when labelling to ensure that the area of the fence board can be accurately marked, especially the edge part and four vertices' points. Moreover, instances of human occlusion, where a human figure temporarily blocks the fence board, present an additional challenge to the labelling process. In these cases, a comparative examination of the frames immediately preceding and following the occluded frame is required to accurately mark the occluded portion of the fence boards. This method allows for a more informed estimate of the obstructed part of the fence board, ensuring the accuracy of the marked key points, even in the presence of obstructions.

Figure 4.3 Example of Annotated Training Data under Various Scenarios

The labelled key points serve as the ground truth for the model. As the Key-point RCNN-PF model is trained, it progressively learns to predict these key points with increasing precision. Every key point that the model accurately predicts represents a step towards the ultimate goal of reliable fence board detection. Despite the laborious nature of this annotation process, it is integral to the model's learning process, supplying it with a reliable and high-quality dataset for effective learning and accurate prediction.

4.1.3 Training and Fine-Tuning the Key-Point RCNN-PF Model

With the dataset ready, the Key-point RCNN-PF model proceeded to be trained. Training a deep learning model is an iterative process that involves feeding the model with the dataset multiple times and adjusting its internal parameters based on the prediction error. These adjustments decrease the discrepancy between the model's predictions and the actual labelled data, refining the model's performance with each epoch. Specifically, the model underwent training across 150 epochs utilizing 105 training images in this case. Notably, this set incorporated 45 images captured from exceptional angles, thus ensuring a diverse training dataset catering to different scenarios the model could potentially encounter.

An essential aspect of the model's training process involved adjusting its learning rate. This parameter determines the extent of modification in the model's internal parameters in response to the observed prediction error in each epoch. After several rounds of comparative testing and evaluation, it was concluded that a learning rate of 0.01 was the most effective. This learning rate facilitated a steady learning process, striking a balance between rapidity and precision, thereby preventing the model from bypassing the optimal solution in the parameter space.

Alongside the learning rate, the weights of the model, integral in determining the contribution of each input feature to the model's output, were meticulously fine-tuned to ensure optimal performance. After training the 150 epochs, the weights corresponding to the 130th epoch produced the highest accuracy levels for fence board detection during subsequent testing phases. The resultant performance indicated a significant improvement in the model's proficiency in detecting fence boards from various angles, reinforcing the efficacy of the chosen training methodology and the model's potentially inaccurate fence board detection.

4.1.4 Results of the Fence Board Detection

4.1.4.1 Results from Test Dataset

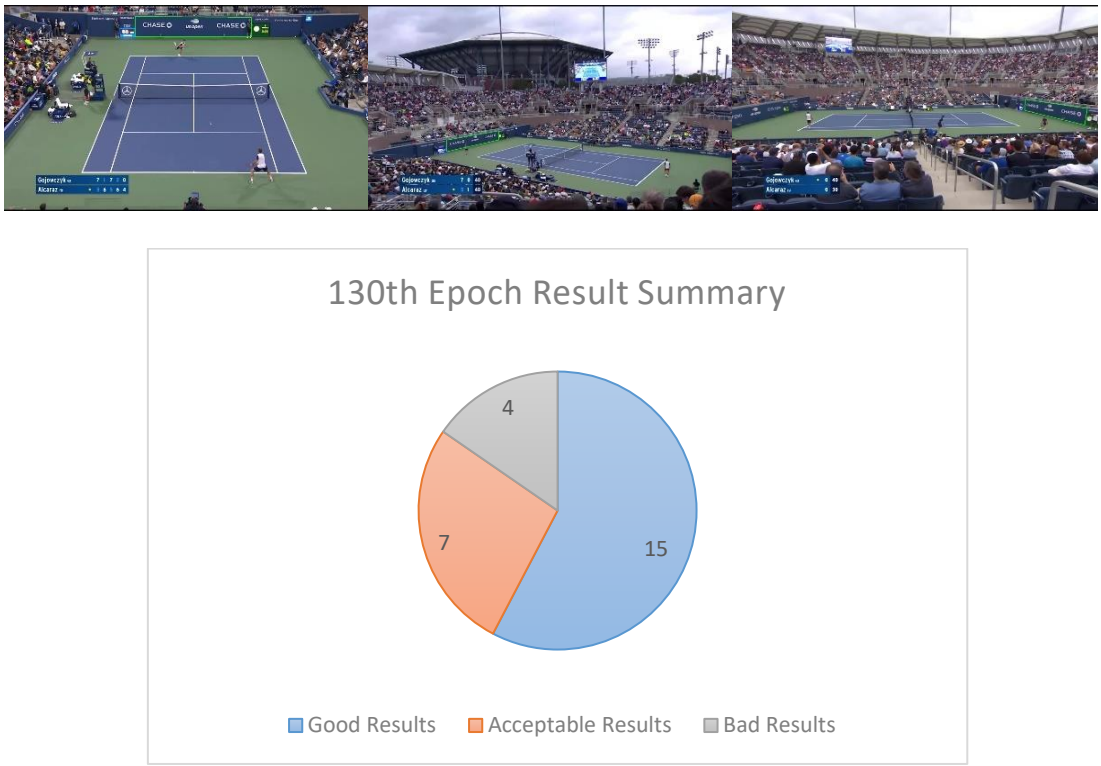

Figure 4.4 Example of Good Results from Test Dataset and Results Summary

The Figure 4.4 above illustrates the outcomes derived from the test dataset corresponding to the 130th epoch, exhibiting instances of successful results. As demonstrated in the leftmost image, despite human occlusion on the right periphery of the fence board, the model has effectively managed to identify and delineate the correct area of the fence board, as indicated by the green rectangle. The middle and right images further substantiate the efficacy of the model's training. Irrespective of the camera angle from which the shots are taken - be it from the left or the right - the model has shown the ability to mark the fence board area accurately. Overall, the results align well with the intended objectives of the model's training, further corroborating the success of the adopted methodology.

Figure 4.5 Example of Bad Results from Test Dataset

The Figure 4.5 above presents an instance where the model's performance didn't meet the expected standards. Despite the model's ability to discern the general location of the fence board within the scene, it falls short in accurately pinpointing the four vertices and demarcating the rectangular area of the fence board, as evident in the image. By comparing with those from successful instances, it becomes apparent that factors such as scene brightness or alterations in the image's colour may have played a role in this subpar performance. It suggests that the model's ability to mark out the fence board area accurately can be potentially influenced by these factors, indicating areas for future refinement and enhancement of the model.

Figure 4.6 Example of Acceptable Results from Test Dataset

The above Figure 4.6 illustrates a scenario of acceptable results. While the model fails to pinpoint and accurately delineate the boundaries of the fence board and the adjacent display boards with absolute precision, the identification outcome remains within a range that can be deemed acceptable.

4.1.4.2 Results from the Demo Video Clip

Figure 4.7 Example of Results from the Demo Video Clip

Although the previously mentioned test results from the 130th epoch produced satisfactory model training outcomes, the application of the pre-trained model to a continuous video demo revealed a somewhat inconsistent performance. This manifested as a failure to recognize and mark the area of the fence board consistently across frames. Consequently, various results of varying degrees of accuracy were obtained, leading to changes in the marked fence board area frame by frame. This inconsistency was represented visually as an incessantly fluctuating rectangular pattern. The Figure 4.7 above display consecutive screenshots taken as the model was applied to the demo video clip. In these images, the fence board area is indicated by a red rectangle. As evident, with the increase in the number of frames, the fence board area within each frame exhibits a different red rectangular shape, or in some instances, the area is not identified at all.

4.1.4.3 Evaluation and Analysis

The findings from the test dataset reveal that despite substantial efforts to enrich the training data, the model still struggles to consistently recognize the fence board area in varying situations within the same video. A potential solution might involve increasing the volume of images in the training dataset. However, considering the considerable time and labour required to label images manually and the subsequent evaluation of each epoch's training results, this task presents significant challenges.

When applying the model to demo video clips, it was observed that the model could often recognize and mark the fence board area. However, due to the dynamic nature of the video - with the fence board area changing in each frame, the camera moving horizontally and vertically, image distortion, and fluctuations in ambient brightness during video recording - the model's performance was inconsistent. For instance, as shown in the Figure 4.8 below, the model could recognize the fence board area and yield what it assessed as high-scoring results. However, these results were unsatisfactory from a human perspective, leading to underwhelming overall outcomes.

Figure 4.8 Example of Unsatisfactory but with High-Scoring Result

Given these limitations, finding solutions to these issues becomes crucial for future research and model improvement. In this thesis, we propose a solution to the above situation and explain it in detail in the subsequent sections.

4.2 Human Occlusion Solution

4.2.1 Data Collection and Annotation

For the human occlusion solution, an accurate and diverse dataset was essential. The data collection phase involved identifying 150 images where humans were in front of the fence boards, including 58 images of players near the edge of the green time display board and blue speed display board, as shown in Figure 4.9. An additional set of 15 images was reserved as the testing dataset to validate the model's performance. The figures presented below illustrate a variety of scenarios: for instance, the close-up image depicts the tennis player in action in front of the fence board or a scenario where the player intersects with the edge of the fence board. Other instances include situations where the ball picker and the tennis player concurrently appear in front of the fence board area or a person overlapping the left or right edge of the fence board. Given the challenges in detecting the fence board within the dataset, efforts were made to accumulate as much data as possible featuring humans appearing in various circumstances, either in front or at the edges of the fence board. This comprehensive approach aimed to tackle potential situations that may arise during the subsequent human segmentation application of the MODNet model.

Figure 4.9 Example Training Data of the MODNet Model

The selected images were annotated using LabelMe, which required the manual identification and marking of the edges of human figures near the fence board, as well as the green time and blue speed display boards. Each human figure in the annotated image was strategically chosen to furnish the model with comprehensive data regarding the shape and location of humans within the frames. The Figure 4.10 below indicate that images that capture long-distance or sideways shots can significantly complicate labelling tennis players or ball pickers. This complexity arises not only due to the diminished size of the human figures in such

parts of the frame but also due to the ensuing blurriness that occurs when these portions of the -image are magnified for detailed examination.

Figure 4.10 Example of Annotated Training Data of the MODNet Model

The ensuing Figure 4.11 present instances where the tennis player and ball picker are situated in front of the fence board or at its edge. The concurrent presence of multiple individuals inevitably escalates the time investment and complexity of the labelling process.

Figure 4.11 Example of Instances of The Concurrent Presence of Multiple Individuals

Given that the tennis player and ball picker frequently appear within the area of the fence board while in motion, the task of enlarging the image for labelling often results in certain human features, such as the head or limbs or sizable portions of the human figure, like the upper or lower half, becoming notably indistinct. This obscurity necessitates repeated viewing of the preceding and following frames during labelling. Such observation aids in discerning the action or posture of the tennis players and ball pickers and facilitates the identification of the blurred elements via comparison of the different frames. This intensive and detailed examination is crucial in accurately differentiating human subjects from the surrounding background elements.

4.2.2 Training and Fine-Tuning MODNet

In the wake of data preparation, the training phase for the MODNet model was initiated. This phase is similar to the training process for the Key-point RCNN model in its iterative nature, which involves feeding the training dataset into the model, making adjustments to the internal parameters contingent upon prediction error, and assessing the model's performance using the testing dataset. The MODNet model was initially trained throughout 200 epochs, with a learning rate set to 0.001 and a batch size of 1. The learning rate and batch size constitute critical

hyperparameters during the model training, shaping the model's capacity to learn from the data. The specific values for these hyperparameters were selected to enable a consistent learning pace while averting any possibility of surpassing the optimal solution.

After an extensive series of iterations and comparative assessments, it was established that the configuration of model weights at the 170th epoch delivered the optimal performance for human segmentation. This was particularly true in segmenting tennis players and ball pickers around the fence board area. Despite the reduced scale of the human figures in fence board areas of some of the frames, the outcome of the human segmentation was sufficiently accurate to address the occlusion issue. Through this repetitive cycle of training, evaluation, and fine-tuning, the MODNet model was adequately equipped to handle human occlusions in the demo videos, thereby enhancing the visual effectiveness of the Ads replacement system.

4.2.3 Results of the Human segmentation

4.2.3.1 Results from test dataset

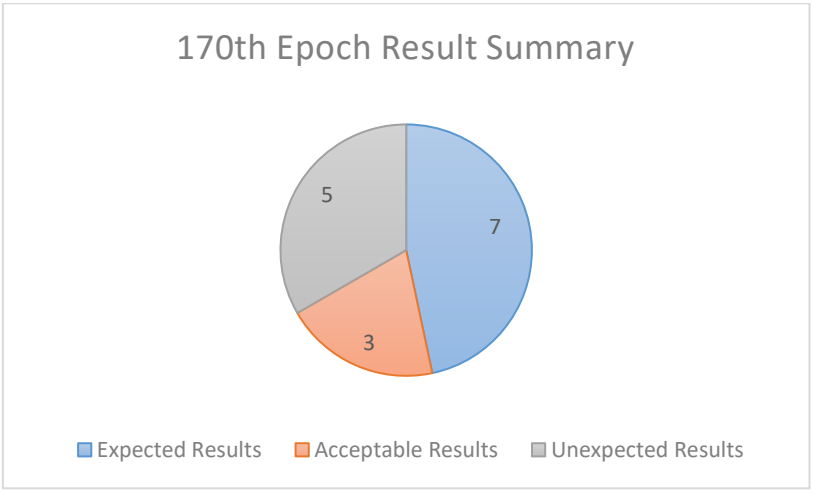

Figure 4.12 Example of Results from Test Dataset and Results Summary

The Figure 4.12 above demonstrate results from the test dataset corresponding to the 170th epoch, showcasing successful outcomes. Active tennis players and ball pickers in front of the fence board have been effectively segmented from the background and highlighted in a distinctive pink hue. Owing to the comprehensive preparation of the training data, effective

segmentation can be achieved in varied scenarios: whether it's a tennis player appearing at the bottom edge of the fence board, ball pickers positioned at the edge of the blue speed display and the green time display, or even the scenario depicted in the middle image where the ball picker traverses the boundaries of both the blue speed display and the fence board. Such segmentation capabilities align well with the desired objectives.

Figure 4.13 Example of Unexpected Results from Test Dataset

While the training yielded the anticipated results, some unexpected outcomes emerged. As can be observed from the provided Figure 4.13, these were due to motion blur occurring when the tennis players or ball pickers were in motion. For instance, when they were running or waving their hands, the positions of their legs and hands exhibited varying degrees of blur, which subsequently impacted the model's segmentation capability. It's also worth noting that lights positioned in the lower region of the fence board were inadvertently segmented alongside the nearby tennis players or ball pickers. This discrepancy is likely attributed to the striking similarity in their colour.

4.2.3.2 Results from the demo video clip

Figure 4.14 Example of Unexpected Results from the Demo Video Clip

When the pre-trained MODNet model was applied to the demo video clip, the human segmentation results largely fulfilled the anticipated objectives. While certain shortcomings were still observable, for example as Figure 4.14, in some frames, the model struggled to effectively segment the white attire of the tennis player, these minor inaccuracies were still considered within a tolerable range. Nonetheless, this underscores the need for ongoing refinement to improve the model's segmentation performance, particularly distinguishing fine-grained features and managing rapid changes in the visual environment.

4.2.3.3 Evaluation and Analysis

Based on the outcomes derived from both the test dataset and the demo video clip, it is evident that the MODNet model, supplemented by a robust and ample training dataset, possesses significant capability in executing human segmentation tasks to a high standard, fulfilling anticipated ideal outcomes. As such, the next phase of this research project involves integrating the MODNet model with the Key-point RCNN-PF model and the Ads pasting functionality. This combined application will be further tested on the demo video clip, providing a more extensive evaluation of MODNet's human body segmentation capabilities. This integration application results will be explained in detail in the subsequent sections.

4.3 Integration of MODNet and Key-point RCNN-PF with Ads

4.3.1 Results of the Integration of MODNet and Key-Point RCNN-PF

Figure 4.15 Example Results of the Integration of MODNet and Key-Point RCNN-PF

The Figure 4.15 above provide a sequential visual representation of screenshots obtained when the integrated model was implemented on the demo video clip. Examining the complete application video, it is clear that the fence board detection by the Key-point RCNN-PF model and human segmentation by the MODNet model achieved performance levels commensurate with previous separate tests. Moreover, the smooth operation of these two models within the integration, without mutual interference, ensures that the visual impact of the entire integrated model's application video meets anticipated expectations.

The integrated model can accurately identify and mark the fence board's area and distinctly isolates the active tennis players and ball pickers within the vicinity of the fence board from the background. This successful outcome bolsters confidence for the subsequent application, which entails the incorporation of the advertisement paste replacement functionality into this integrated model. The advertisement pasting and replacement feature is designed to insert and fill the prepared advertisement image within the red rectangular frame demarcated by the Key-point RCNN-PF model.

From the accompanying images, it's not challenging to discern that the integrated model also carries over the respective limitations intrinsic to both the Key-point RCNN-PF and MODNet models. To illustrate, the top-left and bottom-left images reveal that due to the motion blur of the tennis player and the relative transparency of the tennis racket, coupled with the diminutive size of the tennis ball, the MODNet model struggles to effectively segment the player's blurred hands, tennis racket, and tennis ball. From the top-right image, it becomes apparent that the lighting fixtures near the tennis player at the bottom of the fence board were inadvertently segmented along with the human figure. This likely arises due to colour similarities and spatial proximity, which may confuse the model. In the bottom-right image, one can observe that the Key-point RCNN-PF model cannot precisely delineate the boundaries of the fence board and the blue speed display. Additionally, the MODNet model faces challenges in effectively segmenting the white clothing of the tennis player. These issues likely arise from factors such as rapid movement, light reflections, or subtle colour differences that complicate the process of accurate human segmentation.

These instances underscore some of the persistent challenges faced in achieving precise segmentation and detection in dynamic, real-world scenarios, thereby highlighting areas for potential improvement in future model iterations.

4.3.2 Results of the Integration of MODNet and Key-Point RCNN-PF with Ads

Figure 4.16 Example Results of the Integration of MODNet and Key-Point RCNN-PF with Ads

The Figure 4.16 above present a sequential snapshot of the application video, demonstrating the integration model's advertisement pasting and replacement feature in action. Upon scrutiny of the entire application video, the efficiency of the advertisement pasting process becomes apparent. The advertisement images are seamlessly integrated onto the fence board, while the tennis players are distinctively segmented, contributing to a visually harmonious viewing experience.

The successful fence board replacement by the advertisement maintains the primary focus on the players, ensuring that the background advertisement doesn't detract from the foreground action. This effectiveness can be attributed to the MODNet model's crucial role in human segmentation. Regardless of the background fence board being substituted with an advertisement image, the tennis player remains prominently in view and isn't significantly obstructed. This ensures that the authenticity of the original video is maintained, thereby preserving the audience's viewing experience. It demonstrates that the integrated model can effectively marry the dual objectives of subtle advertising and unobtrusive sports viewing.

Despite the model's overall efficacy, a few wrinkles still need to be ironed out. Notably, in specific frames, the original light fixtures situated at the bottom of the ad are still discernible, as evidenced by the images in the bottom left and bottom centre. This oversight arises from the lamp fixtures being grouped and segmented along with the tennis player. As a result, when an advertisement replaces the fence board, the lamps and tennis player appear juxtaposed against the ad, causing a visual anomaly.

Moreover, as seen in the image on the lower right, the advertisement obscures a significant part of the tennis player's upper body. This is attributable to the limitations inherent to the human segmentation process of the MODNet model. While the model is generally effective, in certain instances, it fails to adequately segment the entire human figure, resulting in these parts being overwritten by the advertisement. This indicates the need to refine the human segmentation component further to ensure an optimal, obstruction-free visual experience.

4.3.3 Evaluation and Analysis

Drawing upon the results generated from the application of the integrated MODNet and Keypoint RCNN-PF models, along with the Ads feature on the demo video clip, it is apparent that specific challenges have arisen. These include issues such as lamp fixtures being mistakenly grouped and segmented with the tennis player or substantial portions of the tennis player's upper body being obstructed by the advertisement. Nevertheless, it's crucial to note that these issues only manifest in a limited number of frames. Given that a standard video comprises 30 frames per second, these anomalies have a minor impact on the overall viewer experience, remaining within acceptable parameters for the duration of this demo video clip.

Though these problems warrant future research for refinement and enhancement, they will not be exhaustively described in the ensuing sections of this article. Instead, the focus will shift to addressing other challenges that significantly influence the viewer experience in the demo video.

4.4 Ads Location Tracking

Figure 4.17 Example Results of FlowNet Applied to the Demo Video Clip

The Figure 4.17 above portrays the outcome generated when FlowNet was applied to the demo video clip. FlowNet's primary function is to estimate optical flow by analysing subsequent frames, which identifies the pixel displacement within the image. A colour-coding scheme for the optical flow field is employed to visualise these flow outcomes. In this scheme, the direction of the vector is represented by colour, and the shade of the colour depicts the magnitude of the vector. For instance, in the above image, white signifies an absence of optical flow, whereas green indicates that the vector of the area is pointing towards the bottom left. A lighter shade of green suggests a relatively minor movement of the camera towards the bottom left. FlowNet is

utilised in this study for the fence board's four vertices. By understanding the vector movement of these vertices in the preceding frame, the positions of the vertices in the current frame can be accurately predicted, if without any other external influences. This facilitates the achievement of Ads location tracking.

FlowNet's application and its results have achieved the expected outcomes. Additionally, it offers a promising solution to the issues identified during the fence board detection process. The specifics of how this is accomplished will be discussed in more depth in the subsequent section on the initial integration model.

4.5 Initial Integration Model

As outlined in the Methodology chapter and preceding sections, the preliminary integrated model amalgamates three separate models - Key-point RCNN-PF, MODNet, and FlowNet. This composite structure offers a unified approach towards addressing key points detection, human occlusion, and Ads location tracking. The FlowNet model integration explicitly presents a resolution for the challenges identified in the fence board detection process. By leveraging the unique capabilities of FlowNet, this integrated framework can effectively estimate and account for frame-to-frame changes, thereby enhancing the overall accuracy and consistency of fence board detection.

4.5.1 Fence Board Detection Issue

As mentioned in the preceding section regarding fence board detection, the performance of the pre-trained Key-point RCNN-PF model exhibited some inconsistency when applied to a continuous video demo. This inconsistency manifested as a lack of uniformity in identifying and demarcating the area of the fence board across different frames. This inconsistent identification led to results with varying levels of accuracy, which, in turn, caused the marked area of the fence board to fluctuate from the previous frame to the current frame. Visually, this inconsistency was observed as a perpetually oscillating rectangular pattern.

To address this challenge, a solution based on FlowNet was proposed. Utilizing FlowNet's capabilities, this solution aimed to ensure a more consistent and stable identification of the fence board area, regardless of the variations in individual frames. The initial integration model begins with the Key-point RCNN-PF model undertaking the task of fence board detection across individual frames. However, due to minor variations such as subtle camera movement or changes in lighting, minor alterations occur in each frame's image, resulting in unique rectangular shapes identified as the fence board in each frame.

In response to this issue, the initial integration model is designed to perform fence board detection at specific intervals—every 10, 30, or 60 frames. To ensure a consistently marked rectangular frame appears on each frame, thus maintaining the advertisement's continuous presence in the fence board's position, the model employs the FlowNet model in the frame verification stages. Under this design, the model only executes keypoints detection on the fence board in the first and every tenth, every thirtieth, or every sixtieth frame of the input video. During the intervals between these frames, the FlowNet model steps in. It calculates the current keypoints based on the previous keypoints detection results, predicts the new keypoints position, and uses these newly predicted positions to provide the necessary rectangular area for advertisement replacement. For instance, after the model identifies key points and sketches the rectangular area for the fence board in the first frame, the FlowNet model assumes the Key-point RCNN-PF model's responsibilities from the second to the tenth frames. It predicts and provides the necessary rectangular area for the advertisement images in this period.

Figure 4.18 Example Results of the Model with FlowNet (Left) and the Model without FlowNet (Right)

The Figure 4.18 above is a comparative illustration of a demo video clip, contrasting scenarios where FlowNet is employed versus instances where it isn't utilized. Upon scrutinizing this relative visual display, it is discernible that the application of FlowNet mitigates the change of advertisement shape across successive frames.

4.5.1.1 Evaluation and Analysis

This solution reduces the frequency of keypoint detections by utilizing the FlowNet model's ability to calculate new keypoint positions based on the inertia of keypoint movement. This diminishes the change of Ads on each frame and curtails the continuous oscillation phenomenon in the visual appearance of advertisement images, ultimately leading to a smoother viewing experience.

Figure 4.19 Example of the Comparison of the Impact of Different Frequencies of Keypoint Detection on Demo Video Clip

The Figure 4.19 above provide a side-by-side comparison of the impact of different frequencies of keypoint detection on a demo video clip. The left portion showcases a model that conducts keypoint detection every 60 frames, while the one on the right does the same operation every 30 frames.

Theoretically, the visual performance of the model performing keypoint detection every 60 frames should surpass that of the one doing it every 30 frames. This expectation is premised on the idea that as the frequency of detections decreases, so does the alteration of the advertisement's shape, thus enhancing the overall viewing experience. This proposition is supported by the video on the left, which exhibits reduced morphological changes in the advertisement and improved continuity and stability of the ad over sequential frames. However, upon a more careful evaluation of the demo video, new challenges came to light. While a reduced frequency of detections undoubtedly enhances the stability of the advertisement's presentation, it does not necessarily guarantee an improved viewing experience for the audience. This is because the visual manifestation of the ad is contingent on the accuracy of the rectangular area provided by the Key-point RCNN-PF model.

For instance, if the model offers an optimal rectangular area during detection, the model performing keypoint detection every 60 frames would display a well-structured advertisement for the subsequent 60 frames or the equivalent of two seconds of the actual video. Nonetheless, if the model produces an irregular, fence-board-like rectangular area, as depicted in the images above, this abnormality will persist for two seconds until the next round of detection by the

Key-point RCNN-PF model. The consequence of this is the continuous display of a distorted ad for a significant duration, which severely impairs the overall performance of the ad replacement system and detracts from the viewer's viewing experience.

To tackle this newly observed problem, we propose a final solution elaborated upon in the subsequent section on the final integration model.

4.5.2 Scene Change Issue

Implementing the initial integration model on the demo video clip brought some compelling results but uncovered fresh challenges. Despite the robust synergy between the Key-point RCNN-PF model and the FlowNet model, which allowed the advertisement to occupy the fence board area reliably, this persistence became a hindrance when a scene change occurred. As the video progresses from one scene to the next, the content displayed can shift dramatically. However, due to the consistent operation of the Key-point RCNN-PF and FlowNet models, the advertisement continues to appear in the previously identified fence board area, even when it no longer exists on the screen. Unless the frame value coincidentally aligns with the 60-frame interval, the initial integration model will not conduct a new fence board detection for the current frame. Consequently, the FlowNet model persistently uses the previously established key point positions for its computations, which are then utilised to assign new positions for the Ads. This peculiar occurrence is referred to as the Scene Change Issue within the context of this thesis.

Figure 4.20 Example of Scene Change Issue (Images order from 1 to 4)

The Figure 4.20 provided above illustrates this problem. As the video frame progressively blurs, it is discernible that the advertisement continues to appear within the previously detected fence board area, even though this area is no longer visible within the new frame. This creates a rather abrupt and aesthetically unpleasing experience, which disrupts the continuity of the original video and impairs the viewer's experience.

Although adjusting the frequency of the fence board detections in the initial integration model (for example, performing a fence board detection every five frames) can potentially mitigate this issue, it simultaneously risks inciting the Fence Board Detection Issue, as mentioned earlier. We propose a novel solution that leverages the scene detection function to address this Scene Change Issue.

4.5.2.1 Scene Detection by PySceneDetect

Handling video data often begins with a critical step: detecting and differentiating the diverse scenes present within the video. This initial task is made considerably manageable with the employment of a robust tool like PySceneDetect. Designed for automatic recognition of scene alterations or cuts in videos, PySceneDetect exploits the variations in colour and intensity between successive frames as a mechanism to discern and pinpoint scene transitions. The provision of precise timestamps corresponding to the inception of each distinct scene acts as a facilitator in segmenting sports game videos into various scenes. This division into discernible scenes promotes more streamlined processing and analysis, making the task significantly efficient. When employed in conjunction with FlowNet, PySceneDetect enhances the uninterrupted tracking of Ads locations across different video scenes. FlowNet aids in estimating the optical flow between frames, providing a coherent vision of the Ads location throughout the different scenes, thereby ensuring a smooth visual transition and maintaining viewer engagement.

```
(pytorch_venv) keane@keane-B660-DS3H-AX-DDR4:~/flownet2-pytorch$ python detect_scene.py
ffmpeg: error while loading shared libraries: libopenh264.so.5: cannot open shared object file: No su
ch file or directory
VideoManager is deprecated and will be removed.
Scene detection finish :)
[0, 678, 766, 832, 904, 1557, 2083, 2163, 2318, 2467, 2731, 2795, 2810, 2825, 2840, 2855, 2870, 2885,
 2917, 2932, 2972, 3246, 3491, 3563, 3670, 3898, 3913, 3931, 4057, 4174, 4206, 4221, 4242, 4284, 4422
, 4701, 4784, 4892, 5035, 5369, 5543, 5574, 5590, 5605, 5620, 5635, 5650, 5665, 5680, 5702, 5748, 576
3, 5778, 6450, 6923, 7570, 8040, 8263, 8786, 8998, 9013, 9028, 9043, 9060, 9088, 9146, 9275, 9359, 93
83, 9398, 9519, 9609, 9735, 10556, 10762, 10914, 11035, 11170, 11644, 11753, 11850, 11871, 11910, 125
16, 12903, 13011, 13068, 13218, 14079, 14123, 14191, 14266, 14359, 14466, 14607, 15147, 15170, 15185,
 15203, 15218, 15233, 15248, 15327, 15342, 15357, 15372, 15397, 15689, 16030, 16163, 16186, 16201, 16
216, 16231, 16246, 16261, 16305, 16483, 16905, 16959, 17000, 17205, 17634, 17697]
```

Figure 4.21 Example Results of the Scene Detection Function

The above Figure 4.21 shows the results obtained when the Scene Detection functionality was applied to the demo video clip. After extensive testing and tweaking, the most

optimal threshold was determined to successfully detect the specific frames where scene transitions occurred within the demo video. The numbers showcased in the figure—0, 678, 766, and so on—represent the frame counts at which these scene changes occur. Incorporating this Scene Detection functionality is invaluable in effectively pinpointing the precise frames where the scenes change. This level of precision in scene transition detection significantly enhances the process of addressing the challenges associated with tracking Ads locations throughout the video. By clearly demarcating the scene transitions, the function aids in maintaining a coherent visual change and continuity in Ads location, contributing to an improved viewer experience.

The scene detection function, which is designed to identify the precise frames where scene changes occur, has necessitated a modification of the criteria used by the initial integration model to invoke the Key-point RCNN-PF and FlowNet models. Upon reaching a frame identified by the scene detection function, the Key-point RCNN-PF model is called upon to perform a timely fence board detection for the current frame. At the same time, the FlowNet model waits in reserve until the Key-point RCNN-PF model identifies new key points.

4.5.2.2 Results and Evaluation

The following Figure 4.22 is illustrative example from a comparison video that alternates between using and not using the scene detection function.

Figure 4.22 Example Result of the Comparison of the Model with Scene Detection Function (Left) and the Model without Scene Detection Function (Right)

Figure 4.23 Example Result of the Comparison of the Model under Scene Change Issue. (Left: With Scene Detection Function and Right: Without Scene Detection Function)

The composite Figure 4.23 above is a visual demonstration of a demo video clip, comparing scenarios in which the scene detection function is utilized versus those where it is not. The left-hand image portrays the model in operation with the enabled scene detection function, while the right-hand image represents the model without the scene detection function. The comparison of the top and bottom images readily reveals that when a scene change occurs, the model depicted on the right (without the scene detection function) persistently displays the advertisement on each frame due to the uninterrupted operation of the FlowNet model. Conversely, the model shown on the left, which utilizes the scene detection function, responds to the onset of a new scene by calling on the initial integration model to engage the Key-point RCNN-PF model on time, effectively suspending the replacement of the fence board area by the advertisement.

Further, when comparing the Figure 4.24 and Figure 4.25 below, it becomes clear that the model using the scene detection function (on the left) continues to deliver superior performance even as the scene changes again. With the support of the scene detection function, the initial integration model is able to summon the Key-point RCNN-PF and FlowNet models swiftly when a new scene materializes. This prompt response ensures accurate detection of the fence board area and enables the model to sustain the ad replacement system seamlessly.

Figure 4.24 Example Result of the Comparison under Scenario that without Perfect Fence Board Area after Scene Change

Figure 4.25 Example Result of the Comparison under Scenario that with Perfect Fence Board Area after Scene Change

Due to the productive implementation of the scene detection function, the Scene Change Issue was successfully mitigated. Building upon this success, the initial integration

model revisited the optimal fence board detection frequency to refine the performance further. The images below juxtapose the effects of differing frequencies of keypoint detection on a demo video clip. The depiction on the left portrays a model that undertakes keypoint detection every 60 frames, whilst the representation on the right conducts the same operation every 30 frames.

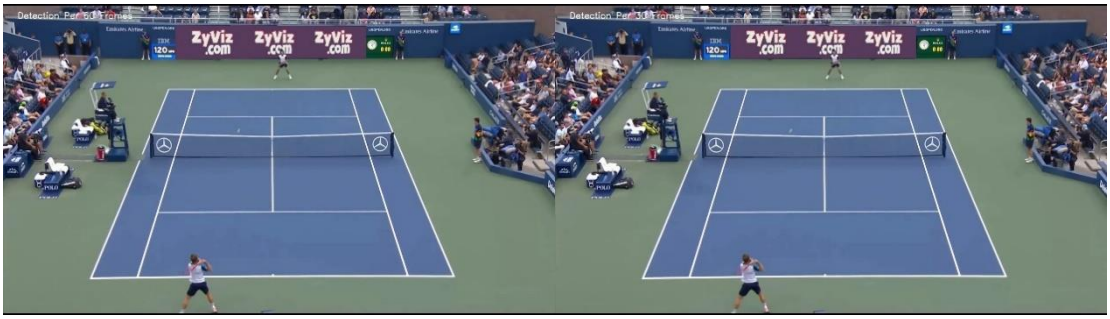

Figure 4.26 Example Result of the Comparison of the Model Operate the Keypoint Detection in Different Frequencies (Left: Operate the Keypoint Detection every 60 Frames and Right: Operate the Keypoint Detection every 30 Frames)

On reassessing the above two settings in light of the supportive role of the scene detection function, it becomes evident that the model which conducts keypoint detection every 60 frames yield superior results. The integration of scene detection appears to balance the frequency of keypoint detection, ultimately leading to a more coherent and stable presentation of the Ads across frames. Thus, the systematic adjustment of detection frequency, in conjunction with scene detection, appears to maximize the performance of the integrated model and effectively address the challenges posed by varying scene changes.

4.6 Final Integration Model

Having successfully overcome the scene change issue, the initial integration model prepared to face the ultimate test: its deployment on a ten-minute demonstration video to gauge its holistic performance and resilience. However, as previously highlighted in Section 4.5.1.1, Evaluation and Analysis of the initial integration model, a newly identified challenge concerning fence board detection surfaced. Specifically, the interval frame number for fence board detection of the initial integration model is fixed at 60 frames, and the model can sometimes produce an irregular, fence-board-like rectangular area. If such an anomaly is produced, it will persist for a full two seconds before the Key-point RCNN-PF model can conduct the next round of detection. This phenomenon, a manifestation of the new issue around fence board detection, magnifies the shortcomings of the initial integration model, particularly when applied to this ten-minute demo video. Moreover, the elongated duration of the demo video resulted in uncovering some previously undetected flaws within the initial integration model.

Upon the completion of the evaluation and analysis phase for the initial integration model's performance with a ten-minute demonstration video, several new challenges have emerged, each demanding attention and resolution:

Fence Board Detection Challenge: The primary issue arose from the constant fluctuation of Ads that replaced the fence boards. This continuous movement created a visually disconcerting effect, preventing the ad from appearing as a static element against the backdrop and disrupting the viewer's immersion.

Advertisement Pasting Challenge: Several sub-challenges arose within this category: Shape Distortion: The first issue was that the Ads, when superimposed on the fence board, did not retain their typical rectangular shape. Instead, they assumed an irregular, rectangular-like shape. This distortion significantly impacted the ad's final presentation, making it appear incongruous within the broader context of the video.

Infinite Extension: The fence board was not wholly visible within the frame in certain shots. This partial visibility resulted in Ads being displayed in an infinitely extended state, which created a visually aberrant result.

Key Point Misalignment: During the process of advertisement replacement, the key points calculated by the FlowNet model sometimes failed to conform to the pre-determined order. This misalignment led to scenarios where the advertisement could not be seamlessly pasted onto the fence board, such as when the upper left corner of the ad was incorrectly placed at the lower left or upper right corners of the fence board.

Human Occlusion Challenge: A problematic interaction between the replaced Ads and the human within the frame in specific frames was observed. Specifically, when the pixels of a human overlapped with the vertex of an ad, the movement of the human would also cause the corresponding vertex of the ad to move. For example, if the lower left corner of an advertisement overlapped with a pixel associated with a tennis player, any movement to the left by the tennis player would result in a corresponding movement of the advertisement's lower left corner. This presented an unnatural interaction that significantly disrupted the advertisement's visual stability.

Recognizing the need for further refinement, we proposed the final integrated model as a solution to address and overcome these newly exposed challenges within the context of this ten-minute demonstration video. This final integration model is expected to solve the emerging issues and enhance the robustness and overall performance of the integrated model in video ad replacement tasks.

4.6.1 Fence Board Detection Challenge

Upon further investigation, it has been ascertained that the primary reason behind the persistent oscillation of the replaced Ads on the fence boards is the minor fluctuation present in each video frame. While these minute fluctuations are typically invisible to the naked eye under standard viewing conditions, their impact becomes starkly noticeable once an advertisement is superimposed onto the fence board. This is mainly due to inconsistencies between the advertisement's fluctuation value or motion vector and each video frame. As a result, when the ad is pasted, it appears to be in continuous motion about the fence board as the video progresses.

As a solution to this challenge, in the final integrated model, the FlowNet model is invoked at the outset to calculate the displacement of the central region across each frame of the video. Following acquiring the central region's displacement data, the mean displacement value is computed and contrasted with the previously obtained displacement data for the advertisement across each frame. This comparative analysis facilitates the determination of an optimal threshold to control the invocation of the Key-point RCNN-PF model for key-point detection.

Due to hardware constraints and the considerable time investment required for such an analysis, the data from the previously applied 30-second demo video clip was employed for this comparative analysis. Upon close inspection of the 30-second demo clip, the 900 frames of the complete video were categorized into two distinct sections: moving frames and stable frames, based on the magnitude of the movement of the central area within each frame. In this thesis, the focus was directed towards the stable frames of the video and the mean displacement value of the four vertices of the advertisement.

Number Of Frames	Mean Values of the AD's Four keypoints displacements
60	0.126921058
120	0.511404634
180	0.004442
240	0.55989
300	3.146634102
660	0.062326733

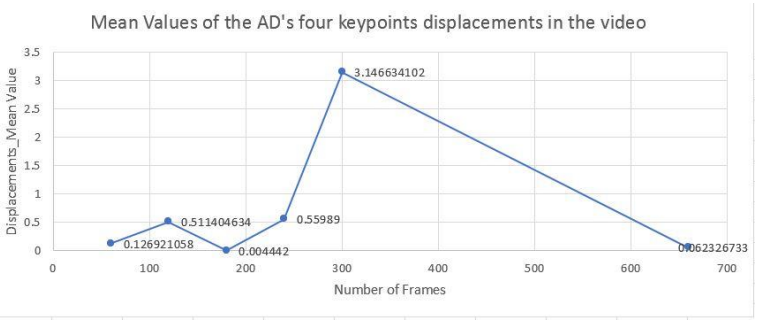

Figure 4.27 Mean Displacement Value of the Four Vertices of the Advertisement

As the model's setting to conduct key point detection every 60 frames, as depicted in the above Figure 4.27, six detections were performed within this 30-second demo video clip. Following a thorough testing cycle, it was concluded that a threshold of 0.126 was optimal for conducting key point detection. This effectively resolved the challenge of the persistent

oscillation of Ads, allowing the replaced advertisement to remain relatively static in relation to the overall video most of the time.

4.6.1.1 Results and Evaluation

The Figure 4.28 below exhibits an exemplary application of the final integrated model on a 30-second demonstration video clip. For this specific application, the model engages in key point detection whenever the mean displacement value of the video's central region attains a value of 0.126. From this 30-second video, it can be observed that during the segments of the video that are stable - specifically from the 5th to the 10th second, and then again from the 25th to the 30th second - the advertisement and the video content maintain a relatively static relationship, which is the anticipated outcome.

Figure 4.28 Example Results of the Final Integrated Model applied on the Demo Video Clip

The analysis intentionally excluded certain types of frames. Among these were moving frames characterised by a substantial range of movement. Despite the FlowNet model's ability to calculate new key point positions, the perceptual experience still had the advertisement jumping erratically from one location to another. The fluctuation was significant enough to disrupt the continuity of the video visually. Additionally, frames without any fence board were also set aside from the investigation. These scenarios had no key points or designated areas for the advertisement. Since the study's primary goal was to analyse how the model performed concerning replacing Ads on a specified 'fence board' area, frames lacking this essential element were deemed irrelevant to the overall evaluation process.

Further exploration discovers that any disparities between the advertisement and the other most of video frames are induced by the Shape Distortion Issue, which is part of the Advertisement Pasting Challenge. Addressing this Shape Distortion Issue will help rectify the perceived inconsistencies between the advertisement and the video frames, thus enhancing the

overall visual continuity and coherence. This impending solution will be discussed in the forthcoming sections, showcasing how we intend to further optimise the model's performance in the context of advertisement replacement.

4.6.2 Advertisement Pasting Challenge

4.6.2.1 Shape Distortion

As shown in the Figure 4.29 below, the lower left segment of the advertisement exhibits distortion, deviating from the expected rectangular shape. This anomaly stems from the inaccuracies in key-point detection by the Key-point RCNN-PF model. Rather than accurately identifying and delineating the fence board area, the model erroneously assigned an atypical region for the ad replacement, resulting in a distorted advertisement display.

Figure 4.29 Example of Shape Distortion

To address this issue, the final integrated model applies further analysis to the four vertices of the advertisement using the FlowNet model. The same 30-second demonstration video clip was utilized to gather relevant data for this process. This resulted in the acquisition of displacement data for all four vertices of the Ads throughout the video. After analysis, it was discovered that the minimum displacement value among the vertices was 0.006.

Using this analytical insight, the final integration model has been further improved. Once the FlowNet model is called upon, it calculates four new key points based on the locations of the previous key points. However, an additional computation layer is performed between the old and the new key points. If the displacement distance between any of the new key points and their corresponding old key point is less than 0.006, the final integrated model interprets this as the new key point being the same as the old one. As such, the FlowNet model refrains from updating the key point's position, opting instead to utilize the old key point location to define

the replacement area for the advertisement. This measure ensures the consistency of the advertisement's shape by preventing unintended distortions in its presentation.

4.6.2.1.1 Results and Evaluation

Figure 4.30 Example Results of the Shape Distortion Solution

As demonstrated in the above image, the problem of shape distortion related to the advertisement has been effectively mitigated. The advertisement now consistently retains its intended rectangular form throughout the entirety of the video. The successful resolution of this issue further augments the efficacy of the solution proposed for the Fence Board Detection challenge discussed in the prior section. Consequently, the advertisement that substitutes the fence board seamlessly integrates with the overall video, resulting in a significantly reduced degree of perceptual incongruity between the ad and the individual frames of the video. This harmonization enhances the visual coherence and overall appeal of the advertisement within the video context.

4.6.2.2 Infinite Extension

Ads can appear in a state of infinite extension in certain specific frames, as depicted in the Figure 4.31 below. For instance, the ad on the left illustrates that it extends beyond the boundaries of the screen. The ad remains affixed on the screen in the previous frame, and the FlowNet model calculates the position for ad replacement in the subsequent frame based on the displacement information of key points from the previous frame. Suppose there is a significant zoom-in alteration between the previous and next frame (such as only the two right vertices of the fence board appearing in the subsequent frame). In that case, the advertisement pasting process encounters this issue. The right vertices of the ad successfully adhere in the next frame. In contrast, the left vertices continue to seek and attempt to adhere to positions previously calculated by the FlowNet model.

Figure 4.31 Example of Infinite Extension

Considering the scenario above, during the operation of the FlowNet model in the final integrated model, certain restrictions are imposed on the positions of the four key points given the total screen range of 1920*1080. Specifically, if the position coordinates of key points are negative, the coordinate defaulted to 0. If the value surpasses 1919, it defaulted to 1919. Similarly, if the value exceeds 1016, it defaulted to 1016. These adjustments ensure the advertisement remains within the intended screen bounds, eliminating instances of infinite extensions of the ad.

4.6.2.3 Key Point Misalignment

When implementing the final integrated model onto the demonstration video clip, an unforeseen issue surfaced during the process of advertisement substitution. Specifically, the key points computed by the FlowNet model were arranged in a sequence inconsistent with the predetermined order, leading to the advertisement being improperly pasted onto the fence board. For instance, the positional arrangement of the top and bottom vertices and the ad's left and right vertices were disordered. As evidenced in the bottom part of the Figure 4.32 below, the advertisement was affixed to the fence board unanticipatedly. That is, the left and right edges of the ad were pasted to the top and bottom edges of the fence board, leading to areas where the ad failed to cover the fence panels themselves.

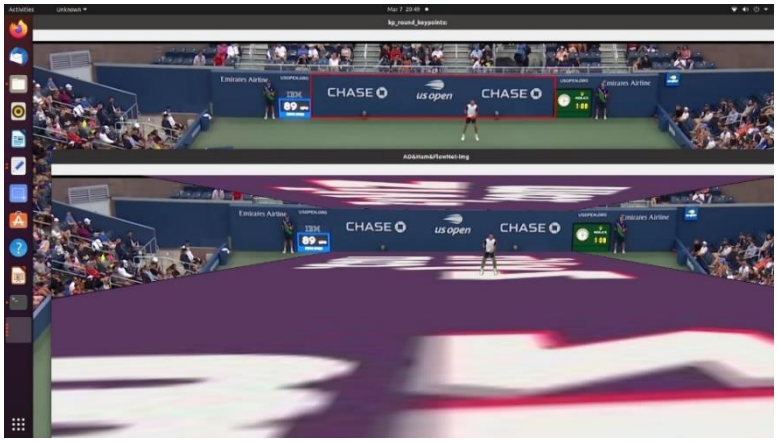

Figure 4.32 Example of Key Point Misalignment

In response to this occurrence, the FlowNet model in the final integrated model has been further optimized and enhanced. When the FlowNet model computes new key points, these calculated new key points are forced to conform to the prearranged sequence, ensuring that the accurate key point positions are obtained before the FlowNet model updates these key point positions. This strategic resolution effectively addressed the Key Point Misalignment Issue and significantly augmented the performance of the final integrated model in terms of advertisement pasting.

4.6.3 Human Occlusion Challenge

During the execution of the final integrated model on the demonstration video clip, a significant interaction issue arose between the replaced Ads and the segmented human pixels. Specifically, when an ad was successfully affixed onto the fence board, the existence of the segmented human pixels overlapping with the bottom left or bottom right corners of the ad presented a challenge. The bottom corners of the ad would follow the movements of the human pixels, resulting in the advertisement experiencing a form of pulling deformation. As illustrated in the Figure 4.33 below, when the tennis player crosses the bottom left corner of the ad, that corner of the ad mimics the left and right movements of the player, almost as though it were affixed to the player himself.

Figure 4.33 Example of Overlap Issue

Following extensive analysis and testing, a resolution was discovered for the Human Occlusion Challenge by applying the Shape Distortion Issue solution to the final integrated model. This solution ensures that when the positions of the four vertices of the ad change, the FlowNet model will execute a displacement calculation between the new and old key points. If the calculated displacement distance is less than 0.006, the model will default the change in key point location to zero. This method effectively solves the issue of the advertisement's distortion due to human movement in the video.

4.6.3.1 Results and Evaluation

Figure 4.34 Example Result of the Comparison of the Model under Human Occlusion Challenge (Left: Without Displacement Distance Calculation Function and Right: With Displacement Distance Calculation Function)

The composite Figure 4.34 above provides a visual representation of a demo video clip, comparing the circumstances where the displacement distance calculation function is deployed against those not implemented. The image's left panel depicts the model's performance sans the displacement distance calculation function. In contrast, the right panel showcases the improved output with the displacement distance calculation function.

As the figure above shows, the Human Occlusion Challenge brought about by the overlap of segmented human pixels has been effectively mitigated. This successful resolution implies that the final integrated model has adeptly tackled all the complications and challenges that initially surfaced with the preliminary integrated model. Furthermore, this also underscores the capability of the final integrated model to proficiently undertake the task of ad replacement proficiently, demonstrating its overall effectiveness and utility in such a complex video processing context.

4.6.4 Results and Evaluation of Final Integration Model

After addressing the problems above and the challenges from analysing the initial integration model's performance, the Final Integration Model is applied to the identical ten-minute demonstration video. This re-application aimed to assess the overall robustness and implementation of the final integration model against the backdrop of the initial model.

It's important to mention that this selected ten-minute demonstration video is extracted from a full-fledged tennis match, which spans approximately three hours. The continuous ten-minute video incorporates numerous scene transitions and predominantly features frames where the tennis player or the ball picker is positioned in front of the fence board area. The rationale behind this selection stems from the video's rigorous testing conditions. Not only does this video offer the opportunity to evaluate how effectively the final model replaces Ads on the

fence board amidst scene changes, but it also provides an avenue to analyse the model's resilience to instances of human occlusion. This is particularly important given the context of a tennis match video, where most frames involve the tennis player or the ball picker moving across the fence board area. This intricate interaction and continuous movement offer an ideal testing ground to validate the final integration model's performance in challenging real-world conditions.

The Figure 4.35 below presents an exemplary application of the Final Integration Model in a ten-minute demonstration video. From this video, it can be decisively inferred that the performance of the Final Integration Model significantly surpasses that of the initial integrated model. The enhanced model is equipped not only to tackle the issue of scene changes effectively but also to proficiently counteract the various problems and challenges that were identified within the framework of the initial integration model. Explicitly addressing the Fence Board Detection Challenge, the Final Integration Model expertly mitigates the issue of continual fluctuations in the Ads substituting the fence boards. In most video frames, the advertisement seamlessly maintains relative stillness in concert with other components on the original screen. This efficient blending of the advertisement into the surrounding visual background prevents the creation of any visually jarring continuous fluctuations, as were previously associated with the advertisement. This precise synchrony and stability present the advertisement as an integral part of the video frame.

Figure 4.35 Example Results of the Final Integrated Model applied a Ten-Minute Demo Video

In the Advertisement Pasting Challenge, the Final Integration Model has proficiently addressed the problems associated with the infinite extension of the ad and the misplacement of key points. Any potential anomalies that could have arisen while pasting the advertisement onto the video have been rectified. The issue of advertisement shape distortion, a primary concern in the initial model, has been significantly mitigated. In most video frames, Ads maintain their typical rectangular shape, thereby preserving the visual consistency of the video

content. However, despite these improvements, a limited number of frames exist where the advertisement shape exhibits partial distortions, as depicted in the Figure 4.36 below.

Figure 4.36 Example of the Final Integrated Model with Partial Shape Distortion Issue

This phenomenon can be attributed to two primary factors. Firstly, the Key-point RCNN-PF model requires extensive training to optimize its performance. The preparation of additional training data can significantly enhance the model's precision in identifying and marking the fence boards. Secondly, certain limitations are inherent to the Key-point RCNN-PF model itself. As with any artificial intelligence model, a 100% recognition accuracy is unachievable due to real-world data's intrinsic uncertainties and variabilities. Therefore, while we can strive to minimize inaccuracies through continuous model refinement, a degree of error tolerance must be factored into the evaluation of model performance.

In the Human Occlusion Challenge, the Final Integration Model has demonstrated significant efficacy in resolving the abnormal interaction triggered by overlapping pixels of Ads and humans in the video. Upon reviewing the entirety of the video, it becomes evident that the Human Occlusion issue has been effectively addressed. Pixels representing segmented humans - such as tennis players or ball pickers - were successfully returned to their original locations following the pasting of the advertisement, ensuring the newly pasted ad did not obscure them. However, upon conducting an in-depth analysis of each frame, some residual problems were identified. During the initial preparation of training data, not enough consideration was given to elements such as tennis balls and rackets that may also occupy the foreground of the fence board area. Consequently, in specific frames of the video, tennis balls in flight or rackets being manoeuvred might be impeded by the advertisement when they traverse over the fence board area.

While such instances are infrequent and occur on a relatively small number of frames, they nonetheless impact the viewer's immersive experience. Any object that unexpectedly disappears behind the advertisement, such as a flying tennis ball or a swiftly moving tennis racket, disrupts the natural continuity of the scene. Therefore, while the Final Integrated Model has made substantial strides in tackling the Human Occlusion challenge, this finding indicates

that further improvements can be made to perfect the ad replacement process and maintain the viewer's sense of immersion.

Figure 4.37 Example Results of the Final Integrated Model Applied to Demo Video Clip with Camera Motion

The Figure 4.37 above captures a moment in the video where the Final Integration Model is applied to a sequence featuring significant camera motion (moving frames). In this fragment, the camera is engaged in a continuous zooming action, moving from a more distanced perspective to a close-up view of the tennis court. This dynamic motion leads to a substantial transformation in the pixel composition of the area due to the fence board region becoming progressively larger in the frame. As a result of this rapid, large-scale enlargement of the fence board area, the advertisement's ability to adhere tightly to the fence board is compromised. Instead of maintaining a snug fit throughout the zooming motion, the ad begins to drift, developing a noticeable gap from the fence board.

This issue becomes particularly evident when the camera reaches a closer, more intimate perspective. In these frames, a significant discrepancy is observed between the advertisement's upper edge and the fence's top. This disjuncture creates a perceptible separation between the ad and the fence board, diminishing the desired seamless integration of the advertisement within the video sequence. Thus, while the Final Integrated Model handles static and mildly dynamic scenes proficiently, it exhibits challenges in maintaining consistency during scenes featuring intensive camera motion, such as zooming.

The Final Integration Model significantly surpasses its predecessor, the initial integration model, regarding capability and precision. While it is acknowledged that some minor deficiencies remain, it is important to recognize that the model has achieved the

anticipated benchmarks and successfully navigated a series of demanding tasks. In essence, while there is always room for further enhancement, the Final Integration Model has accomplished its task commendably, demonstrating robust performance across various challenges.

4.6.5 New Application of Final Integration Model

Considering the successful demonstration of the Final Integration Model's capabilities, an exploration was undertaken to apply this model to effect advertisement replacement in a new area. This undertaking signifies a novel usage of the Final Integration Model, broadening its application beyond the confines of its initially delineated scope, the fence board area, to encompass the new, previously uncharted area. This attempt highlights the model's flexibility, bringing its adaptability and resilience to the test across various visual environments. It underscores the model's capacity to tackle various scenarios, emphasizing its potential for future development and deployment across a broad spectrum of visual recognition tasks. The Figure 4.38 below showcases the environment where the application of the advertisement replacement is attempted. The red rectangle delineates the designated placement area for the advertisement, symbolizing the targeted surface where the ad is intended to be affixed. Concurrently, the green rectangle represents the detection area, signifying the field within which the model actively identifies suitable advertisement placement.

Figure 4.38 Example of Annotated Training Data in New Region

Following a methodology to the previously employed data processing and model training regimen, the Final Integration Model was applied to paste Ads within the newly designated region. The Figure 4.39 below provides a representative example of a video clip showcasing this application of the Final Integration Model.

Figure 4.39 Example Results of the Final Integrated Model Applied to Paste Ads into New Region

Upon inspection of the resultant video, it can be discerned that the Final Integration Model adeptly navigates this new environment. It accomplishes the task of pasting Ads within this new region, maintaining the visual integrity of the Ads and ensuring they harmoniously blend with the surrounding area. The pasted Ads appear relatively stationary, mirroring the static nature of the area. Importantly, they are seamlessly integrated into the visual landscape, causing no discernible visual disturbance or fluctuations. This outcome underscores the model's flexibility and ability to adapt and deliver under varying circumstances.

Naturally, given that the Final Integration Model has not been specifically adapted or further refined to account for this new application, certain imperfections manifest themselves. These deficiencies, as depicted in the Figure 4.40 below, include challenges related to the precision of area detection and issues with the distortion of the pasted advertisement's shape.

Figure 4.40 Other Example Results of the Final Integrated Model Applied to Paste Ads into New Region

Although with a few problems, this is nonetheless a significant and valuable exercise. It demonstrates the model's potential and underscores its adaptability in handling diverse operational scenarios. More importantly, it highlights the areas requiring attention and offers a roadmap for future enhancement. These minor setbacks essentially delineate the contours of the model's potential development space and hint at the future directions for its refinement.

Chapter 5: Conclusion and Future Work

5.1 Conclusion

In the present thesis, a solution is successfully proposed grounded in the domain of Computer Vision (CV) that facilitates the replacement of virtual advertisements Ads on fence boards in sports game videos. This achievement was realized through a step-by-step exploration and analysis of three distinctive models: the Key-point RCNN - Picture Frame (Key-point RCNN-PF) model, the Matting Object Decomposition Network (MODNet) model, and the FlowNet model. Each of these models was tested using specifically curated training and test data sets to scrutinize their respective performances in the face of unique task scenarios. To elaborate, the capability of the Key-point RCNN-PF model was evaluated for fence board detection. In the case of the MODNet model, its assessed efficacy in handling human occlusion solutions, while the performance of the FlowNet model was evaluated in the context of tracking the location of Ads.

Building upon the foundation established by these three models, an initial integrated model was conceptualized. This prototype model was then applied to a continuous ten-minute video clip excerpted from a real-life tennis match. This practical application evaluated the direction of the proposed solution and provided crucial insights that led to the conception of the Final Integration Model. This iterative process of evaluation, analysis, and enhancement ultimately underpins our innovative solution for the replacement of virtual Ads in sports game videos. Following a thorough cycle of adjustments, optimization efforts, and enhancements to the integration among the three foundational models within the initial integration model, the Final Integration Model was obtained and eventually arrived at refined. This culminating model offers the following achievements and contributions to the domain of Virtual Advertisement (Ad) Replacement on Fence Boards within Sports Game Videos, as detailed below:

In the fence board detection challenge, the Key-point RCNN-PF model was employed to successfully identify and annotate the fence board area within sports game videos. Through a coordinated application of both the Key-point RCNN-PF and FlowNet models, the visually problematic issue of the fluctuating placement of the advertisement replacing the fence board was effectively addressed. Consequently, throughout most of the frames in the demonstration video, the ad maintains a level of stability congruent with other elements present within the

original image. The result is a seamless blend into the backdrop, effectively eliminating the previously observed visual fluctuations.

Addressing the advertisement pasting challenge, the integration of the advertisement achieved success within the video frame's fence board area. Utilizing the FlowNet model, the ad's positioning was effectively tracked, and in tandem with PySceneDetect, the scene change issue was resolved. This ensured the advertisement would appropriately respond to changes in the scene. For instance, the ad would not be displayed without the fence board in the video frame. Simultaneously, various issues related to the pasting of Ads were resolved, such as the dilemma of indefinite ad extension and the misplacement of key points during the ad replacement process. This ensured the ads were correctly and aesthetically integrated into the changing scenes, greatly enhancing the viewing experience.

Additionally, the issue of advertisement shape distortion has been significantly mitigated, indirectly optimizing the results of fence board detection. Although this solution does not directly enhance the model's detection accuracy, it ensures that the identified fence panel area is depicted in its customary rectangular form in most instances.

Regarding the human occlusion challenge, the desired outcome was effectively accomplished. Specifically, after the replacement of the advertisement on the fence board, the pixels corresponding to humans commonly seen in proximity to the fence board - for instance, the tennis player or ball picker - will not be overshadowed by the advertisement. This ensures that there is no interference between the representation of these individuals and the newly integrated advertisement, thereby preserving the naturalness and fluidity of the scene and enhancing the viewing experience.

5.2 Recommendation for Future Work

In the process of testing the Final Integration Model, a series of shortcomings and constraints were identified. Specifically, in the fence board detection challenge, due to a lack of training data, the accuracy of fence board detection did not meet the anticipated performance benchmarks set for the Key-point RCNN-PF model. Similarly, in the human occlusion challenge, while the model was capable of successfully segmenting human pixels present in front of the fence board area, it faltered when confronted with certain specific frames, particularly those involving tennis balls flying over the fence board or tennis rackets moving across the fence board area. This lack of effective segmentation in these instances resulted in a compromised viewing experience.

Given these observations concerning the Final Integration Model and its performance in the new application, future improvements can be envisaged along the following lines:

- Procuring a larger dataset for training the Final Integration Model, which could help enhance its performance in fence board detection and human segmentation tasks.
- Collecting additional data encompassing moving frames and optimizing model parameters to augment the model's effectiveness when processing fragments involving motion.
- Exploring and experimenting with other detection and segmentation models to ascertain whether they could contribute to bolstering the overall performance of the Final Integration Model in the realm of advertisement replacement.

References

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/s0731-7085(99)00272-1

Aggarwal, A., Mittal, M., & Mittal, M. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004. https://doi.org/10.1016/j.jjimei.2020.100004

Almgren, K., Krishnan, M., Aljanobi, F., & Lee, J. (2018). AD or Non-AD: A Deep Learning Approach to Detect Ads from Magazines. *Entropy*, *20*(12), 982. https://doi.org/10.3390/e20120982

Astolfi, G., Rezende, F., De Andrade Porto, J. V., Matsubara, E. T., & Pistori, H. (2021). Syntactic Pattern Recognition in Computer Vision. *ACM Computing Surveys*, *54*(3), 1–35. https://doi.org/10.1145/3447241

Bara, B. a. H., Pokrovskaia, N. N., Ababkova, M. Y., Brusakova, I. A., & Korban, A. A. (2022). Artificial Intelligence for Advertising and Media: Machine Learning and Neural Networks. In *2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. https://doi.org/10.1109/elconrus54750.2022.9755590

Barron, J. A., Fleet, D. J., Beauchemin, S. S., & Burkitt, T. (2003). *Performance of optical flow techniques*. https://doi.org/10.1109/cvpr.1992.223269

Black, M. J., Yacoob, Y., Jepson, A. D., & Fleet, D. J. (2002). *Learning parameterized models of image motion*. https://doi.org/10.1109/cvpr.1997.609381

Black, M. M., & Anandan, P. (2002). *A framework for the robust estimation of optical flow*. https://doi.org/10.1109/iccv.1993.378214

Boda, J., & Pandya, D. (2018). *A survey on image matting techniques*. https://doi.org/10.1109/iccsp.2018.8523834

Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *Lecture Notes in Computer Science* (pp. 25–36). https://doi.org/10.1007/978-3-540-24673-2_3

Brox, T., & Malik, J. (2011). Large displacement Optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(3), 500–513. https://doi.org/10.1109/tpami.2010.143

Byon, K. K., & Phua, J. (2021). Digital and Interactive Marketing Communications in Sports. *Journal of Interactive Advertising*, *21*(2), 75–78. https://doi.org/10.1080/15252019.2021.1970422

Carmigniani, J. C., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., & Ivkovic, M. (2011). Augmented reality technologies, systems and applications. *Multimedia Tools and Applications*, *51*(1), 341–377. https://doi.org/10.1007/s11042-010-0660-6

Chai, J., Zeng, H., Li, A., & Ngai, E. W. (2021). Deep learning in CV: A critical review of emerging techniques and application scenarios. *Machine Learning With Applications*, *6*, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

Cheng, W., Song, S., Chen, C., Hidayati, S. C., & Liu, J. (2021). Fashion Meets Computer Vision. *ACM Computing Surveys*, *54*(4), 1–41. https://doi.org/10.1145/3447239

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: object detection via region-based fully convolutional networks. In *arXiv (Cornell University)* (Vol. 29, pp. 379–387). Cornell University. https://arxiv.org/pdf/1605.06409

Ding, X., Li, Q., Cheng, Y., Wang, J., Bian, W., & Jie, B. (2020). Local keypoint-based Faster R-CNN. *Applied Intelligence*, *50*(10), 3007–3022. https://doi.org/10.1007/s10489-020-01665-9

Dosovitskiy, A., Fischery, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., & Brox, T. (2015a). *FlowNet: Learning Optical Flow with Convolutional Networks*. https://doi.org/10.1109/iccv.2015.316

Dosovitskiy, A., Fischery, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., & Brox, T. (2015b). *FlowNet: Learning Optical Flow with Convolutional Networks*. https://doi.org/10.1109/iccv.2015.316

Du, Z., Liu, J., & Wang, T. (2022). Augmented Reality Marketing: A Systematic Literature Review and an Agenda for Future Inquiry. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.925963

Efimova, V., Fedotov, L. E., Shalamov, V., & Filchenkov, A. (2022). *Advertisement replacement in video*. https://doi.org/10.1117/12.2623581

Esteva, A., Chou, K. J., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E. J., Dean, J., & Socher, R. (2021). Deep learning-enabled medical CV. *Npj Digital Medicine*, *4*(1). https://doi.org/10.1038/s41746-020-00376-2

Ghassab, V. K., Maanicshah, K., Bouguila, N., & Green, P. E. (2020). *REP-Model: A deep learning framework for replacing ad billboards in soccer videos*. https://doi.org/10.1109/ism.2020.00032

Giombi, K., Viator, C., Hoover, J., Tzeng, J., Sullivan, H., O'Donoghue, A. C., Southwell, B., & Kahwati, L. C. (2022). The impact of interactive advertising on consumer engagement, recall, and understanding: A scoping systematic review for informing regulatory science. *PLOS ONE*, *17*(2), e0263339. https://doi.org/10.1371/journal.pone.0263339

Girshick, R. (2015). *Fast R-CNN*. https://doi.org/10.1109/iccv.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation*. https://doi.org/10.1109/cvpr.2014.81

Gupta, A., Anpalagan, A., Guan, L., & Khwaja, A. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, *10*, 100057. https://doi.org/10.1016/j.array.2021.100057

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask R-CNN*. https://doi.org/10.1109/iccv.2017.322

Hinton, G. E., & Salakhutdinov, R. (2012). A Better Way to Pretrain Deep Boltzmann Machines. In *Neural Information Processing Systems* (Vol. 25, pp. 2447–2455). https://papers.nips.cc/paper/4610-a-better-way-to-pretrain-deep-boltzmann-machines.pdf

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1–3), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2

Hosang, J., Benenson, R., Dollár, P., & Schiele, B. (2016). What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(4), 814–830. https://doi.org/10.1109/tpami.2015.2465908

Hossari, M., Dev, S., Nicholson, M., McCabe, K., Nautiyal, A., Conran, C., Tang, J., Xu, W., & Pitié, F. (2018). ADNET: a deep network for detecting adverts. *AICS*, 45–53. http://ceur-ws.org/Vol-2259/aics_6.pdf

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I. S., Wojna, Z., Song, Y., Guadarrama, S., & Murphy, K. (2017). *Speed/Accuracy Trade-Offs for modern convolutional object detectors*. https://doi.org/10.1109/cvpr.2017.351

Huang, L., Liu, X., Wang, X., Li, J., & Tan, B. (2023). Deep learning methods in image matting: a survey. *Applied Sciences*, *13*(11), 6512. https://doi.org/10.3390/app13116512

Hui, T., Tang, X., & Loy, C. C. (2018). *LiteFlowNet: a lightweight convolutional neural network for optical flow estimation*. https://doi.org/10.1109/cvpr.2018.00936

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). *FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks*. https://doi.org/10.1109/cvpr.2017.179

Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Jinzhu, L., Tan, L., & Jiang, H. (2021). Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification. *Agriculture*, *11*(8), 707. https://doi.org/10.3390/agriculture11080707

Kalaiarasi, P., & Rani, P. E. (2020). A Comparative Analysis of AlexNet and GoogLeNet with a Simple DCNN for Face Recognition. In *Advances in intelligent systems and computing* (pp. 655–668). Springer Nature. https://doi.org/10.1007/978-981-15-5029-4_54

Kandaswamy, C., Silva, L. R., Alexandre, L. A., Sousa, R. M., Santos, J. E., & De Sá, J. M. (2014). *Improving transfer learning accuracy by reusing Stacked Denoising Autoencoders*. https://doi.org/10.1109/smc.2014.6974107

Ke, Z., Sun, J., Li, K., Yan, Q., & Lau, R. W. H. (2022a). MODNET: Real-Time Trimap-Free Portrait Matting via objective Decomposition. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *36*(1), 1140–1147. https://doi.org/10.1609/aaai.v36i1.19999

Ke, Z., Sun, J., Li, K., Yan, Q., & Lau, R. W. H. (2022b). MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *36*(1), 1140–1147. https://doi.org/10.1609/aaai.v36i1.19999

Ke, Z., Sun, J., Li, K., Yan, Q., & Lau, R. W. H. (2022c). MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *36*(1), 1140–1147. https://doi.org/10.1609/aaai.v36i1.19999

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems* (Vol. 25, pp. 1097–1105). http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D. J., Howard, R. A., Hubbard, W. B., & Jackel, L. D. (1989). Backpropagation applied to handwritten Zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lee, K., & Moon, N. (2021). Digital Signage System Based on IntelligentRecommendation Model in Edge Environment:The Case of Unmanned Store. *Journal of Information Processing Systems*, *17*(3), 599–614. http://jips-k.org/digital-library/2021/17/3/599

Li, X., Li, J., & Lu, H. (2019). A Survey on Natural Image Matting With Closed-Form Solutions. *IEEE Access*, *7*, 136658–136675. https://doi.org/10.1109/access.2019.2941832

Li, Y., Wan, K., Yan, X., & Xu, C. (2005). *Real time advertisement insertion in baseball video based on advertisement effect*. https://doi.org/10.1145/1101149.1101221

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/tnnls.2021.3084827

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). *Feature pyramid networks for object detection*. https://doi.org/10.1109/cvpr.2017.106

Linna, M., Kannala, J., & Rahtu, E. (2018). *Real-time Human Pose Estimation with Convolutional Neural Networks*. https://doi.org/10.5220/0006624403350342

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. M., Fu, C., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In *Lecture Notes in Computer Science* (pp. 21–37). https://doi.org/10.1007/978-3-319-46448-0_2

Liu, Z. (2023). *A Deep Neural Framework To Detect Individual Advertisement (Ad) From Videos*. https://openaccess.thecvf.com/content/WACV2023/html/Liu_A_Deep_Neural_Framework_To_Detect_Individual_Advertisement_Ad_From_WACV_2023_paper.html

Longman, K. A. (1997). If not effective frequency, then what? *Journal of Advertising Research*, *37*(4), 44–50. https://dialnet.unirioja.es/servlet/articulo?codigo=432001
Ma, H., Wan, J., Li, B., & Xia, T. (2017). *Multi-view 3D Object Detection Network for Autonomous Driving*. https://doi.org/10.1109/cvpr.2017.691
Mohamed, A., Hinton, G. E., & Penn, G. (2012). *Understanding how Deep Belief Networks perform acoustic modelling*. https://doi.org/10.1109/icassp.2012.6288863
Morrison-Smith, S., & Ruiz, J. (2020). Challenges and barriers in virtual teams: a literature review. *SN Applied Sciences*, *2*(6). https://doi.org/10.1007/s42452-020-2801-5
Negahdaripour, S., & Yu, C. X. (2002). *A generalized brightness change model for computing optical flow*. https://doi.org/10.1109/iccv.1993.378241
O'Shea, K., & Nash, R. R. (2015). An Introduction to Convolutional Neural Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1511.08458
Pieters, R., & Bijmolt, T. H. A. (1997). Consumer Memory for Television Advertising: A Field study of duration, serial position, and Competition Effects. *Journal of Consumer Research*, *23*(4), 362. https://doi.org/10.1086/209489
Pyun, D. Y., Han, J., & Ha, J. H. (2004). Attitudes and effectiveness toward/of virtual advertising on Major League Baseball. *In Annual Conference for the Sport Marketing Association, November, Memphis, TN*.
Ranjan, A., & Black, M. J. (2017). *Optical flow estimation using a spatial pyramid network*. https://doi.org/10.1109/cvpr.2017.291
Rehman, M. Z. U., Shafiq, M., & Ullah, F. (2022). Automated Computer Vision-Based Construction Progress Monitoring: A Systematic Review. *Buildings*, *12*(7), 1037. https://doi.org/10.3390/buildings12071037
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In *Neural Information Processing Systems* (Vol. 28, pp. 91–99). https://papers.nips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf
Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031
Sander, M., & Altobelli, C. F. (2010). Impact of virtual advertising in sports events. *International Journal of Sport Management and Marketing*. https://kops.uni-konstanz.de/handle/123456789/12728
Sander, M., & Altobelli, C. F. (2011). Virtual advertising in sports events: does it really work? *International Journal of Sports Marketing & Sponsorship*, *12*(3), 28–42. https://doi.org/10.1108/ijsms-12-03-2011-b004
Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1
Shah, S. a. A., & Xuezhi, X. (2021). Traditional and modern strategies for optical flow: an investigation. *SN Applied Sciences*, *3*(3), 289. https://doi.org/10.1007/s42452-021-04227-x
Shah, S. a. A., Xuezhi, X., & Ahmed, W. (2021). Optical Flow Estimation with Convolutional Neural Nets. *Pattern Recognition and Image Analysis*, *31*(4), 656–670. https://doi.org/10.1134/s1054661821040210
Shrivastava, A., Gupta, A., & Girshick, R. (2016). *Training Region-Based Object Detectors with Online Hard Example Mining*. https://doi.org/10.1109/cvpr.2016.89
Simoncelli, E. P., Adelson, E. H., & Heeger, D. J. (2002). *Probability distributions of optical flow*. https://doi.org/10.1109/cvpr.1991.139707
Sood, S., Singh, H., Malarvel, M., & Ahuja, R. (2021). Significance and Limitations of Deep Neural Networks for Image Classification and Object Detection. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. https://doi.org/10.1109/icosec51865.2021.9591759

Sun, D., Yang, X., Liu, M., & Kautz, J. (2018). *PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume*. https://doi.org/10.1109/cvpr.2018.00931
Tellis, G. J. (1997). Effective frequency: One exposure or three factors? *Journal of Advertising Research*, *37*(4), 75–80. https://ci.nii.ac.jp/naid/10021880655
Tu, Z., Xie, W., Zhang, D., Poppe, R., Veltkamp, R. C., Li, B., & Yuan, J. (2019). A survey of variational and CNN-based optical flow techniques. *Signal Processing-image Communication*, *72*, 9–24. https://doi.org/10.1016/j.image.2018.12.002
Turley, L. W., & Shannon, J. R. (2000). The impact and effectiveness of Ads in a sports arena. *Journal of Services Marketing*, *14*(4), 323–336. https://doi.org/10.1108/08876040010334547
Uijlings, J., Van De Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5
Vasudevan, V., & Gounder, M. S. (2021). Advances in Sports Video Summarization – A Review Based on Cricket Videos. In *Springer eBooks* (pp. 347–359). https://doi.org/10.1007/978-3-030-79463-7_29
Vaz, J. M., & Balaji, S. (2021). Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics. *Molecular Diversity*, *25*(3), 1569–1584. https://doi.org/10.1007/s11030-021-10225-3
Viktoratos, I., & Tsadiras, A. (2021). Personalized Advertising Computational Techniques: A Systematic Literature Review, Findings, and a Design Framework. *Information*, *12*(11), 480. https://doi.org/10.3390/info12110480
Voulodimos, A., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, *2018*, 1–13. https://doi.org/10.1155/2018/7068349
Wan, K., & Xu, C. (2006). *Automatic Content Placement in Sports Highlights*. https://doi.org/10.1109/icme.2006.262925
Wang, J., & Cohen, M. B. (2007). Image and video matting: a survey. *Foundations and Trends in Computer Graphics and Vision*, *3*(2), 97–175. https://doi.org/10.1561/0600000019
Wedel, A., Cremers, D., Pock, T., & Bischof, H. (2009). *Structure- and motion-adaptive regularization for high accuracy optic flow*. https://doi.org/10.1109/iccv.2009.5459375
Xie, H., Hou, K., Jiang, D., & Ma, W. (2023). Bust Portraits Matting Based on Improved U-Net. *Electronics*, *12*(6), 1378. https://doi.org/10.3390/electronics12061378
Yang, C., Yu, X., Kumar, A., Ali, G. G. M. N., Chong, P. H. J., & Lam, P. T. (2021). Social Video Advertisement Replacement and its Evaluation in Convolutional Neural Networks. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, *20*(1), 117–136. https://doi.org/10.5565/rev/elcvia.1347
Yao, G. (2017). A survey on Pre-Processing in Image Matting. *Journal of Computer Science and Technology*, *32*(1), 122–138. https://doi.org/10.1007/s11390-017-1709-z
Zhang, W., Fu, C., & Zhu, M. (2020, August 2). *Joint Object Contour Points and Semantics for Instance Segmentation*. arXiv.org. https://arxiv.org/abs/2008.00460
Zhang, X., Wang, G., Chen, C., Dong, H., & Shao, M. (2022). Semi-Supervised Portrait Matting via the Collaboration of Teacher–Student Network and Adaptive Strategies. *Electronics*, *11*(24), 4080. https://doi.org/10.3390/electronics11244080
Zheng, F., Yang, C., Chong, P. H. J., Wang, G. T., Ali, S. N., & Lam, P. T. (2021a). *Deep Learning Algorithm for Picture Frame Detection on Social Media Videos*. https://doi.org/10.1109/iotais53735.2021.9628865
Zheng, F., Yang, C., Chong, P. H. J., Wang, G. T., Ali, S. N., & Lam, P. T. (2021b). *Deep Learning Algorithm for Picture Frame Detection on Social Media Videos*. https://doi.org/10.1109/iotais53735.2021.9628865
ZhiQiang, W., & Jun, L. (2017). *A review of object detection based on convolutional neural network*. https://doi.org/10.23919/chicc.2017.8029130

Zhou, Q., & Yu, C. (2022). Point RCNN: An Angle-Free Framework for Rotated Object Detection. *Remote Sensing*, *14*(11), 2605. https://doi.org/10.3390/rs14112605

Appendices

Demo Video:

The Final Integration Model was applied to New AD Pasting Region Demo (Stable Camera Motion）.avi

The Final Integration Model was applied to New AD Pasting Region Demo (Moving Camera).avi

The Final Integration Model was applied to Fence Board Region under Stable Camera Motion-10min Demo Video.avi

The Final Integration Model was applied to Fence Board Region under Moving Camera Motion-5min Demo Video.avi