Deep Learning Based Advertisement Replacement on Dynamic Background Videos

Cheng Yang [1,*], Fucheng Zheng [1], Duaa Zuhair Al-Hamid [1], Peter Han Joo Chong [1], and Patrick Lam [2]

[1] Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland, New Zealand
[2] Zyetric Technologies Ltd., Hong Kong, China
Email: ych2tj@gmail.com (C.Y.); fucheng.zheng@aut.ac.nz (F.Z.); duaa.alhamid@aut.ac.nz (D.Z.A.-H.); peter.chong@aut.ac.nz (P.H.J.C.); patrick.lam@zyetric.com (P.L.)
*Corresponding author

***Abstract*—Advertising (AD) on video backgrounds reduces the disturbing of audiences. However, integrating virtual AD on videos' dynamic backgrounds is difficult. The object detection method cannot follow the background moving correctly, so that the AD replacement shows flicking on the video. Furthermore, the target background object may move behind the virtual logo and foreground objects. The virtual AD should be inserted behind these things. To overcome those challenges, this research integrates multiple Deep Neural Networks (DNN). First, object detection DNN identified object locations. Thus, tracking of these locations was done through an optical flow DNN. Moreover, an image in-painting DNN reconstructed the blocked objects, which helps the detection approach. Based on the detection and tracking, the virtual AD is pasted on the video, then object segmentation is utilized to put foreground objects back on top of the virtual AD. The experimental results show that, in the dynamic background scenario, the AD replacement has a sensitivity of approximately 81.1%, a specificity of at least 92.57% and a successful rate of more than 83.3%. This means that, in most cases, the virtual AD can be integrated into the appropriate position on the moving background.**

***Keywords*—deep learning, neural networks cooperation, product replacement, dynamic background detection, background inpainting, foreground segmentation**

I. Introduction

Given the ubiquity of multimedia videos, advertising and marketing organisations are interested in providing targeted Advertisements (AD) to clients via online videos. This has a significant impact on both marketing agencies and end users/consumers. The popularity of online videos has provided more avenues for the propaganda of brand information to the audience. Traditionally, virtual ADs can be incorporated into online videos. They are known as product replacements, where specific brands or products are inserted into the online videos [1]. The current computer vision techniques can automatically replace a static object in a video with a virtual AD. However, it is challenging to embed a virtual AD into an object that moves on the video background.

In this paper, we propose a novel dynamic background AD replacement framework that automatically replaces an object on a dynamic background with a virtual AD in a video. This automatic process will significantly help in product replacement for online video advertising. Our approach combines different Deep Convolutional Neural Network (DCNN) techniques to achieve its goal.

There are different types of research on advertising replacement. Jiang *et al.* [2] employed Faster-RCNN to detect unauthorised advertising billboards. Morera *et al.* [3] employed SSD [4] and YOLOv3 [5] to detect outdoor AD panels for replacement. However, the scope of this study is limited to static AD billboard detection. Also, the AD replacement on moving objects in a dynamic background cannot be performed.

Furthermore, some industry projects can do the AD replacement. Their performance is very good in live streaming. The principle is to create the 3D virtual box space based on the camera's intrinsic and extrinsic parameters. These projects require high costs and are only useful on live streaming. However, an online video doesn't provide the camera parameters, so it is impossible to do AD replacement with those projects.

Our proposed method (or model) focuses on dynamic background AD replacement in sports videos. This method doesn't rely on camera parameters. Thus, it is suitable for any 2D sports video. Generally, dynamic background AD replacement approaches encounter four difficulties. First, in a 2D video, the target objects are moving in a 3D background. This decreases the detection accuracy of DCNN. Second, the object moves following the background. It is required that the virtual AD embedding should accurately follow the background movement. Third, the foreground objects, such as athletes, occlude the target objects. Therefore, it is necessary to introduce a virtual AD behind these objects. Finally, virtual logos, such as TV logos and scores, significantly block the target objects, affecting detection accuracy.

The contribution of this research is to identify the corresponding techniques to solve the four problems (tasks) listed above. In addition, it introduces a combination

Manuscript received May 9, 2024; revised May 17, 2024; accepted July 29, 2024; published January 27, 2025.

doi: 10.18178/joig.13.1.34-45

model for overcoming different neural network limitations to make AD replacement possible on a dynamic background. Our research utilizes a Federation International de Football Association (FIFA) soccer game video as an example to show the performance of our proposed deep learning algorithms for AD replacement in videos. The dynamic background objects are the target banners on the fence of the soccer playground. This research aims to replace the existing target banners with a new virtual AD. In this paper, four neural networks have been identified, implemented and combined:

- First, the Keypoint Regional Convolutional Neural Network for Picture Frame (R-CNN-PF) and parallel criterion [6] were used to detect banners. This method could detect the polygon shapes of the banners, reducing 3D angle detection to a 2D detection task.
- Second, Flownet v2 [7] was applied to track the target banner's speed against the moving background.
- Third, MODNet [8] was implemented to segment soccer players, which could resolve the occlusion problem.
- Fourth, a background in-painting method Aggregated Contextual Transformations Generative Adversarial Network (AOT-GAN) [9] was utilized to reconstruct the banners blocked by the virtual logos in the soccer game video.
- Finally, a combination approach is proposed to handle the limitation of the four neural networks, so that they can cooperate well for the aim of AD replacement.

In the end, the monitored target banners were replaced with a virtual AD, demonstrating the performance of the method.

The rest of the paper is structured as follows. In Section II, the related work is discussed. In Section III, the automatic product replacement approach is introduced. Section IV describes the data collection; presents the results of the experiments; and reports the discussion and future directions. Finally, Section V concludes the paper with future work.

II. Literature Review

To the best of our knowledge, the AD replacement on video dynamic background has not been released yet. In one of the few AD replacement studies, Jiang *et al.* [2] used Faster-Regional Convolutional Neural Network (RCNN) to detect illegal billboard advertisements. The detection algorithm employed the multi-angle suggestions area to accurately locate and mark illegal billboards. Morera *et al.* [3] used Single Shot Detector (SSD) [4] and YOLOv3 [5] to detect outdoor AD panels for replacement. In their experiments, both SSD and YOLO detectors produced acceptable results under a variety of panel sizes, illumination conditions, viewing perspectives, partial panel occlusion, complex backgrounds and, multiple panels in scenes. Although these studies were able to detect the advertising billboards, the AD replacement on moving objects with dynamic background was not achieved.

There is a study on AD banner detection in soccer match videos wherein the Hue value was analysed to find the boundary between banners and green playground [10]. The Hough transform was then utilized to detect the edge and find the Regions of Interest (ROI) of the banners. Finally, the target banners are identified by comparing pre-defined template banner images. However, this method had limitations, such as the playground edge being unclear, vertical poles and insufficient contrast.

In summary, the methods discussed above may detect static banners in a video. These detection methods can be used to replace a banner with a virtual AD. However, they cannot be used on a moving banner that is in sync with the background. For example, although a virtual AD can be placed on the (banner) detection site frame by frame, the virtual AD will be flicking since detection does not precisely track the background movement pace. Therefore, the dynamic background product replacement task cannot be fulfilled with a single neural network. It requires various technologies, such as object detection, background moving recognition, foreground segmentation and background in-painting. These technologies should be merged to achieve the final goal.

The first important task in product replacement is object detection, which identifies the target object for AD replacement. Herein, Deep Convolutional Neural Networks (DCNN) can detect objects automatically. Some DCNN technologies may identify object details, such as positions, angles and shapes. Particularly, Iqbal *et al.* [11] could do image classification with a very small dataset. They analysed eight learning algorithms and found an optimal method for neural networks' learning. After that, they researched 24 experiments to find a method for choosing optimum hyperparameters. This method improved image classification and detection a lot. Following this research, Mask R-CNN [12] provides a human Key-point detection feature that can recognize human poses. Zhang *et al.* [13] improved the Mask R-CNN for estimating multi-person postures. Also, they presented a multi-task deep network that is aware of the occlusion part. Westermann *et al.* [14] combined the DCNN and SIFT to detect billboard AD on the soccer game playground fence. Their detection could suit the AD's 3D angle on the video. Zheng *et al.* [6] presented a Keypoint R-CNN-PF to detect four corners of picture frames on social media video. Their detection does not require 3D angle information, but it can detect picture frames from different angles in 2D videos.

The second significant task is to measure background movement. Herein, a pasted virtual AD should move in accordance with the background motion. The Optical flow is an accurate technique for measuring background moving speed. For optical flow measurement, several methods use DCNN. Flownet v2 [7] is the most widely used approach. This approach developed optical flow estimation for a CNN learning problem. It includes three contributions: an optimal training data schedule, a stacked architecture that incorporates warping the second frame with the intermediate optical flow, and a subnetwork specializing in small motions that is used to expound on small displacement. Zhao *et al.* [15] reported that the occlusion area is a difficulty in optical flow measurement. The authors presented an asymmetric occlusion-aware feature-matching module for measuring optical flow. Jiang *et al.* [16] introduced another approach to address the occlusion problem and achieve optical flow. They reported a global motion aggregation module that enhanced optical flow estimates in the occluded regions while preserving performance in non-occluded regions.

The third important task in product replacement is human segmentation. If a target object is in the background, the foreground, such as a human body, may occlude the object. The foreground segmentation can help in obtaining the occluded object’s pixels. DCNN technologies have improved accuracy and speed over the last five years, including Full Convolutional Networks (FCN) [17], Mask R-CNN [12], and U-Net [18]. Furthermore, research like [19] could segment the human body, especially the knee joint synovial fluid. In their neural network, a 9-layer structure for simple feature extraction, followed by 30 layers for complex pattern detection. Then, 75 layers of convolution looked for more detailed features. Finally, after another 9 layers, the output of the segmentation gives precise results. In the current years, image matting is the most recent technique for precisely segmenting objects. The DCNN for image matting consists of tri-map based methods, such as Guided Contextual Attention Image Matting [20] and HDMatt [21], as well as tri-map free methods, such as Refinement Network Matting [22] and MODNet [8].

Furthermore, if foreground objects block the background target objects (used for AD replacement), the object may not be detected accurately, thus, the background of the video needs to be reconstructed. Herein, video inpainting technology is required. The DCNN for image inpainting frequently uses Generative Adversarial Network (GAN) architecture [23]. Some video inpainting methods can work with low resolutions (about 432×240 or even lower.). Liu *et al.* [24] described a FuseFormer method that can tackle the issue of blurry edges in detail. Li *et al.* [25] introduced Flow-Guided Video Inpainting (E2FGVI) to reconstruct the background of a video. This can improve the efficiency and effectiveness of the inpainting process. Some video inpainting supports a resolution of 512×512. Xu *et al.* [26] presented another concept for the flow-guided video inpainting method wherein a spatially and temporally coherent optical flow field across video frames was provided. Herein, the inpainting quality is improved by hard flow example mining. Zeng *et al.* [9] presented an aggregated contextual-transformation generative adversarial network (AOT-GAN) model for video inpainting. They constructed the generator by stacking multiple layers of a proposed AOT block, which can capture both informative distant image contexts and rich patterns of interest for context reasoning. They additionally improved the discriminator by training it with a custom mask-prediction task. Furthermore, unlike existing video inpainting methods, this network is designed for high-resolution image inpainting.

In summary, considering the robustness and easy of implementation, four methods were chosen for the proposed AD replacement model. The Keypoint R-CNN-PF was chosen for target banners detection. The FlownetV2 was selected to measure the Optical Flow. The MODNet is utilized for the foreground segmentation. The background in-painting task was completed by AOT-GAN.

III. Materials and Methods

Based on the challenges introduced in Section I, we present a dynamic background AD replacement model that includes four deep neural networks (DNN): Keypoint R-CNN-PF, Flownet, MODNet and AOT-GAN. Each DNN is implemented for each task. In this section, the four tasks with the corresponding deep neural networks are first introduced. Each task section indicates the limitations and solutions to explore the cooperation and combination of the four neural networks. Finally, the entire combination model is displayed to show the working process of AD replacement.

A. Target Banners Detection and AD Replacement

The soccer field playground’s barrier is covered in banners. They move in the background of the video as the camera moves to follow the ball. The banners’ shapes should be accurately detected so that the virtual AD image can cover the entire area of the detected banners. The Keypoint R-CNN-PF for picture frame detection [6] is utilized to detect the banners in each frame of the video. It can detect the four corner points of a quadrilateral object.

The architecture of the Keypoint R-CNN includes the backbone of Wide-ResNet-50, which is to extract features with different scales. After the backbone, the Region Proposal Networks is to propose anchors. These anchors are for classification and localization refinement. After that, the ROI Align crops and aligns the objects’ ROI for Keypoint detection. Finally, the Picture Frame (PF) Head with Feature Pyramid Network (FPN) architecture extracts the corner points of the banners.

Since the Keypoint R-CNN-PF runs in every frame of the video, it works with banners that move in the background. On the other hand, some banners detection results have incorrect shapes, because the Keypoint R-CNN-PF is not 100% accurate. Thus, the parallel criteria [6] is used to eliminate incorrect shape detection. Therefore, not all banners are detected.

The Keypoint R-CNN-PF can detect all banners in the soccer game video. In this research, only a few target banners would be detected for AD replacement. Other non-targeted banners would not be detected. Target banners can be manually cropped out of a video frame. These cropped images serve as inference images for comparing banners with Keypoint R-CNN-PF (See Section IV.B). Each reference banner image is compared to all detected banners on each video frame, as shown in Algorithm 1.

In Algorithm 1, the content comparison is used to compare the difference between the reference banner image (ref_image) and the detected banner image (det_banner). Each pixel from the reference banner image is subtracted from each pixel in the detected banner image. The mean value of all pixel differences will then be used to calculate the content difference between two images. If the difference is less than a certain threshold, the target banner is determined for the virtual AD replacement.

```
Algorithm 1: Target banners detection.
for ref_image in target_banner_images:
    for det_banner in detected_banners:
        Convert ref_image to HSV colour space
        Convert det_banner to HSV colour space
        Difference = mean(ref_image - det_banners)
        If difference < 43:
            Choose this banner's detection
        else:
            delete this banner's detection
    end
end
```

To find the optimal threshold, the first 10 minutes of the test videos are used for a simple experiment. (The test video is introduced in Section IV.) The content comparison compares these detected banners to the reference banner images. If a detected target banner is correctly determined by the reference banners, this is a true positive. Otherwise, it is a false negative. Furthermore, if a non-target banner is determined as a target banner, this is a false positive. If the threshold is too high, there will be many false positives. Conversely, a very low threshold leads to many false negatives. The detection and content comparison are applied on the 10mins video with different thresholds. The results are virtually observed to see the number of true positives, false negatives and false positives. According to the virtual observation, the optimal threshold in this research is 43, which performs the lowest false negatives and false positives. After the content comparison method, the detected non-targeted banners will be removed. Fig. 1 shows the detected target banners with green boxes.

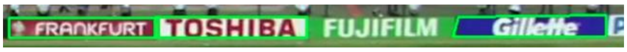


Fig. 1. The target banners detection example. The green quadrilaterals on the banners indicate the detection result. The banners of "Frankfurt", "Toshiba" and "Gillette" are chosen target banners.

When a target banner is detected, the pixel coordinates in its four corners are output. These coordinates are used by the OpenCV function of "cv.getPerspectiveTransform()". This function converts the AD image's four corners to those of the detected target banner. Then, the OpenCV function of cv.warpPerspective() applies the transform to paste the AD photo to the target banners' location. Fig. 2 shows the outcome of the AD replacement, in which the original ADs of "Frankfurt", "Toshiba" and "Gillette" in the banners are replaced by the virtual AD of "ZyViz".

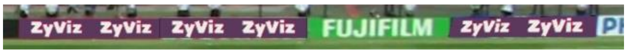


Fig. 2. The outcome of the AD replacement result.

If a target banner is on the video's edge, the AD should be pasted partially. In that case, the two corner points outside the video need to be located. These two corner points can be calculated using banner detection. Suppose the four corners detection on the banners are x_n and y_n, where $n \in [1,4]$. The (x_3, y_3) and (x_4, y_4) are at the edge of the video as shown in Fig. 3. However, the AD photo should be pasted on the two corner points outside the video which are (x_3', y_3') and (x_4', y_4'). For the size ratio, width:height, of the banner to be 7.5:1, after calculating the banner's height, h, using the left two corner points, the new banner's width, w', is:

$$\boldsymbol{w' = 7.5 \times \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} \tag{1}$$

To find the two points outside the video, the slopes on the top and bottom lines need to be used to calculate the two positions of the points as shown in Fig. 3. Take the top line as an example, the slope, Θ, is:

$$\boldsymbol{\theta = \tan^{-1} \frac{y_3 - y_1}{x_3 - x_1}} \tag{2}$$

Then, the point (x_3', y_3') on Fig. 3 is calculated by:

$$\boldsymbol{x_3' = x_1 + w' \times \cos\theta} \tag{3}$$

$$\boldsymbol{y_3' = y_1 + w' \times \sin\theta} \tag{4}$$

The point, (x_4', y_4'), uses a similar calculation from the bottom line. After that, four points (x_1, y_1), (x_2, y_2), (x_3', y_3') and (x_4', y_4') are obtained to use for pasting the AD.

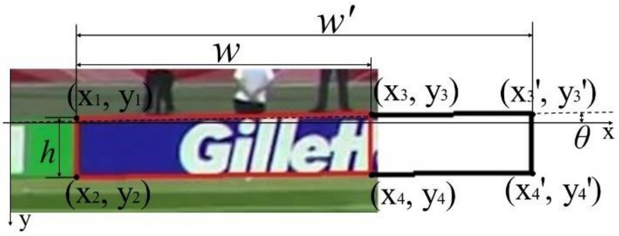


Fig. 3. The geometry diagram for calculating the points outside the video.

After the above calculation, the virtual AD can be pasted following the new four points. An example is shown in Fig. 4 under the next section.

B. Banners Tracking and Scene Change Detection

The Keypoint R-CNN-PF may not detect banners accurately in every frame of the video. Some target banners may be lost. The lost target banners have no detection results. Therefore, the detection method may not detect target banners in all frames. When this problem happens, the detection from the previous frame can be used to locate the banners. However, the banners move with the background. The previous detection cannot be immediately applied. Therefore, an object tracking method should be used to predict (or adjust for) the detection of a non-detected target banner. Because the detection outputs four corners of a banner, in this research, Flownet v2 [7] is utilized to track the target banner's movement.

The Flownet architecture first has 9 convolutional layers followed by max pooling. Following that Flownet v2 includes a stacked architecture which warps the two continuous frames with intermediate optical flow. The final structure is a subnetwork specializing in small motions which can elaborate on small displacement.

The Flownet v2 uses the previous frame's detection (four corners of the banner) to get the related optical flow (or speeds between continuous frames). The speeds are used to calculate the current coordinates of the four corners. For example, the detection method gives a target banner's four corner points (pixel coordinates) from the previous frame. If the current frame does not have the target banner's detection, Flownet is used to find the four points' speeds. Then, the four points' coordinates can be calculated as below:

$$\boldsymbol{x_n^c = x_n^p + v_x^n} \tag{5}$$

$$\boldsymbol{y_n^c = y_n^p + v_y^n} \tag{6}$$

where, x_n^c and y_n^c, $n \in [1,4]$, are the (x, y) coordinates of the target banner on the current frame. x_n^p and y_n^p are the (x, y) coordinates from the previous frame. v_x^n and v_y^n are the (x, y) direction speeds for the four corner points of the target banner. Fig. 4 shows an example of tracking after virtual AD paste.

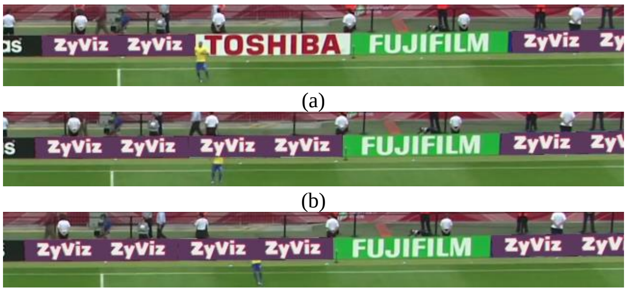


Fig. 4. The AD paste result through several continue frames. (a) The 1st frame. (b) The 10th frame. (c) The 20th frame.

The Flownet v2 is not 100% accurate, therefore, it is applied from the 2nd frame to the 9th frame. Then, detection is triggered in the 10th frame. Following that, the Flownet v2 tracks the target banners' corners from the 11th frame. The process occurs every ten frames. If the detection does not detect the target banners in the 10th frame, the undetected banners are tracked by Flownet v2 for the next 10 frames.

The soccer match has two types of camera views (or scenes): long-shot scene and short-shot scene. The long-shot scene is from the main camera. It shows the whole soccer field. The soccer players are displayed small in this scene. The short-shot scene shows a close view of some soccer players, audiences, and special shortcuts. In our study, the AD replacement only works on the long-shot scene because the short-shot scene may not show the banners. Therefore, if the camera scene changes from a long-shot to short-shot, the banners tracking should be disabled. When the camera scene changes from a short-shot scene to a long-shot scene, the banners detection/tracking and AD replacement should work again.

To find the scene change, a method is introduced to compare the content of two continuous frames. The two video frame images are converted to HSV colour space. Then, each colour channel of hue, saturation and value is compared to the content. Finally, the mean value of the three colour channels' differences is used to check whether it is scene-changing or not. Scene change detection is practiced by using a Python package PySceneDetect.

The PySceneDetect cannot recognize the long-shot or short-shot scenes, but it detects the scene change. A special progress is applied to integrate the scene change detection into the target banner detection and Flownet. Once a video frame is identified as a scene change, the target banners detection/tracking is disabled, and all the tracked banners' information is deleted. After that, the target banner detection is run frame by frame. If the next scene is a short-shot scene, the target banner detection will not detect anything and the tracking will not be executed. If the next scene is a long-shot scene, the target banners can be detected. Also, the tracking with Flownet will be executed in the following frames. Again, the detection and tracking process continues every ten frames until the next scene change is identified.

Another problem is that if a soccer player (human) blocks one of the target banners' tracking corners, the Flownet v2 can only detect the soccer player's optical flow (pixel speed). Herein, the corner tracks may be lost. The solution is that if one corner's optical flow (speed) is much faster or slower than the other three corners', this lost corner's speed is fixed by the average value of the other three corners' speed values. The lost corner's current position is calculated using the prior position and the fixed speed. This is the key point fixed method.

C. Blocked Banners Reconstructing

Because banners are in the background, they can be blocked (even 70%) by the virtual logos on the video. Virtual logos include team information, time, scores and the TV logo shown in the video. If the virtual logos occlude the banners, it is challenging to detect the banners, and the virtual AD should be placed behind the virtual logos. The result of the blocked banner detection is shown in Fig. 5.

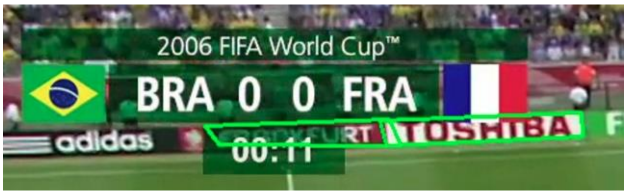


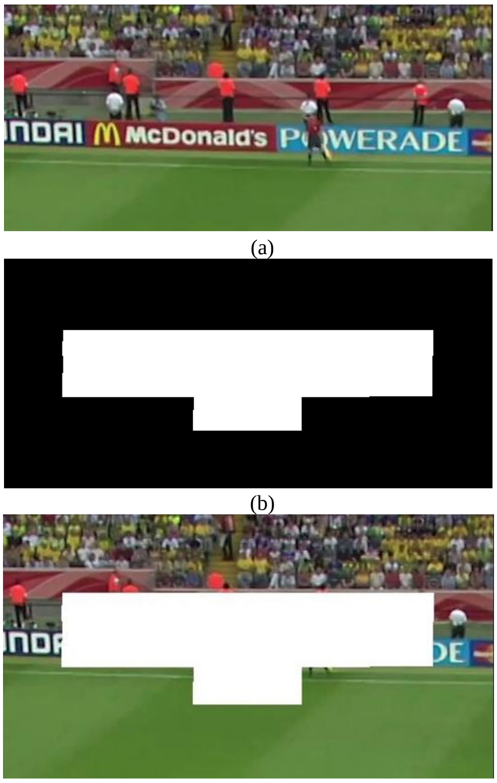
Fig. 5. The result of the blocked banner detection. The banner of "Frankfurt" is blocked by the virtual image. The banner cannot be detected well.

In Fig. 5, the banner, "Frankfurt", is blocked by the virtual image. It cannot be detected well by Keypoint R-CNN-PF. Thus, the quadrilateral shape 'in green' is abnormal. The solution to this problem is to reconstruct the blocked banner, and then use Keypoint R-CNN-PF to detect the banner on the reconstructed image. The reconstructing method is the AOT-GAN [9]. This neural network is good for high-resolution image inpainting.

The AOT-GAN is a GAN-based model for high-resolution image inpainting. The GAN contains a generator and discriminator for generating new content from image patterns. Besides the GAN architecture, it includes AOT blocks which aggregate contextual transformations from various receptive fields, allowing to capture of both informative distant image contexts and rich patterns of interest for context reasoning. This leads to AOT-GAN re-generating banners more accurately.

As the virtual logos are always in the same position in the soccer match video, they can be manually labelled from the first frame and produced the mask. This mask is used for all video frames.

The data to train the AOT-GAN is produced from the virtual logo mask and the original banner images from the soccer video. Fig. 6 shows an example of data for training. The image in Fig. 6 (a) is the original image which is cropped from the video frame. Fig. 6 (b) is the virtual logo mask which is manually labelled. Fig. 6 (c) shows the broken image, which is used to train AOT-GAN.


(a)

(b)

(c)

Fig. 6. The clips of the banner's data for training of AOT-GEN network. (a) Original image. (b) The virtual logo mask. (c) The broken image.

After training, the AOT-GAN can reconstruct the blocked banners on the video implementation as shown in Fig. 7. Although the reconstruction does not recover the banner letters well, the boundary of the banner is clear enough for detection. Some reconstructed images are collected and added to the Keypoint R-CNN-PF training dataset with other non-blocked banner image data.

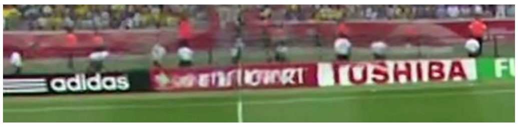
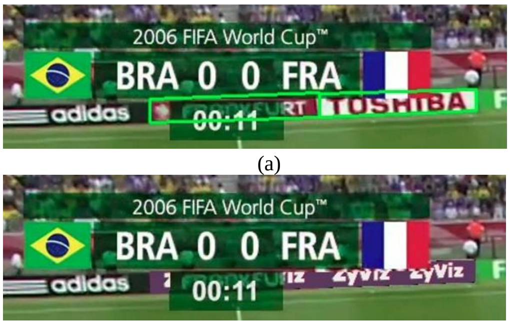


Fig. 7. The result of image in-painting. The blocked banner contents of "Frankfurt" and "Toshiba" are re-constructed.

After the banners reconstruction, the Keypoint R-CNN-PF and content comparison method (from section III.A) detects the banners on the reconstructed image. The result is shown in Fig. 8(a). The banner detection correctly finds the banner's four corners. With the correct detection, the virtual AD can be pasted properly. After that, the virtual logo mask is also used to collect the virtual logo pixels and paste them back onto the virtual AD. Fig. 8(b) shows the AD replacement result.


(a)

(b)

Fig. 8. The AD replacement result after image in-painting. (a) The correct banner detection. (b) The AD replacement result.

Because the AOT-GAN costs computation time, it would not be run for every video frame. It is triggered only if a banner detection or tracking overlaps with the virtual logo region. Therefore, overlap (blocking) recognition is executed after the banners tracking.

D. Human Being Segmentation and Repasting

Based on the sections above, a virtual AD can replace the target banners in the background. However, if a soccer player stands (occludes) in front of a target banner, the pasted AD overlaps the soccer player. The correct AD replacement requires placing the virtual AD behind the soccer player.

The solution is to segment the soccer players and extract the pixels from the video. The segmentation task is done by MODNet [8], which is a neural network for human matting. The MODNet is a three-branches architecture. A low-resolution branch is built by using MobileNetV2. This is for semantic estimation. A high-resolution branch consists of fewer convolutional layers which are input high-resolution images. This is for detailed prediction. At last, a semantic-detail Fusion branch combines the semantics and detail features for final human segmentation. The MODNet can segment small size of human beings, which is suitable for the situation of soccer match videos.

The data for training of MODNet is collected by labelling soccer players who block banners. Fig. 9 shows a label for a football player and its corresponding mask. The mask pixels have special values. The pixels on the player's body are "1". Background pixels are "0". The pixels on the boundary (two pixels thick) between the soccer player and the background are "0.5". The boundary is used to trigger the high-resolution branch.

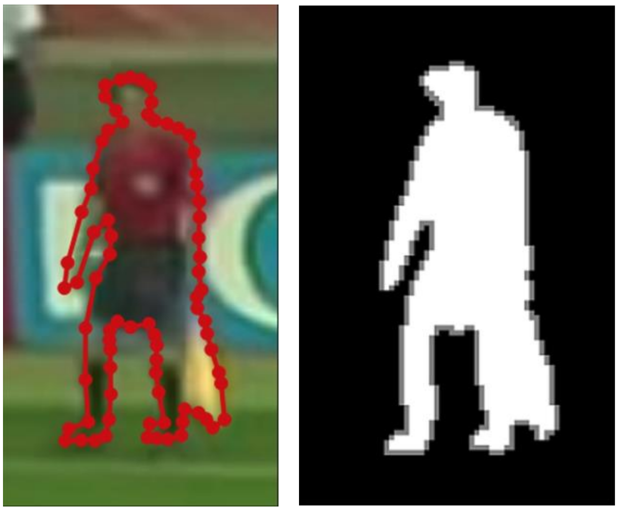


Fig. 9. The example of football player labelling. (a) A football player label. (b) The corresponding mask

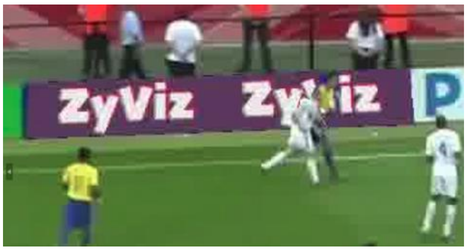


Fig. 10. The example of the football player occlusion on AD replacement.

After the soccer player segmentation, the football players' pixels can be extracted from the original video frame. The virtual AD is then put onto the target banners. After that, soccer player pixels can be placed in the original location and on top of the virtual AD. This is the solution to the human occlusion problem as shown in Fig. 10.

E. Combination Model Process

The entire combination model incorporates deep natural networks and traditional computer vision techniques. The architecture is shown in Fig. 11. Using the football game video as an example, the model aims to replace the target banners' content with a customized virtual AD photo. (In this case, the virtual AD photo is "ZyViz" advertisement, which is shown in Fig. 2).

At the beginning, a video frame is checked whether there is a scene change or not. If the frame is identified as a scene change, the banners detection and tracking information are deleted. Otherwise, all the information is stored. If this is the first frame of a new scene or frame for the detection (11th frame, 21st frame, etc.), the video frame is fed into the Keypont R-CNN-PF network to detect all the banners, and then the content comparison filters out the target banner detection information (banners identities (IDs) and four corner points). This is the target banners detection process.

It is supposed that, if a video frame is a short-shot scene, the detection is empty. The tracking of banners' information is cleared. Then, there is no more operation, the proposed model executes the next frame for target banners detection. The frame count keeps zero. The target banner detection executes every frame until it identifies target banners in a frame (probably a long-shot scene). This is because the proposed combination model supposes the target banners only appear in the long-shot scene.

If the video frame is in a long-shot scene, the target banners detection recognizes the target banners. Then, this information is checked with the previous frame's information to see whether a banner is continually detected, lost or a new banner. (Note: if the video frame is the first frame of a long-shot scene, all target banners are new).

In the next step, all the banner detection results are checked whether they are in the video edge or not. If a target banner is in the video edge, the video edge calculation is triggered to get the new four corners' coordinates. After that, all target banners are replaced by the virtual AD.

If a banner on the current video frame is lost detection, or the video frame is for tracking (frame between 10th, such as 2~9th frame, 11~19th frame, etc.), The Flownet is triggered, and the optical flow (speed) of the four corners of each banner is measured. The banner's position in the current frame is calculated by Eqs. (5) and (6).

Next, the current location of the banners are checked whether they overlap with the virtual logos or not. If there is an overlap, the video frame is fed into AOT-GAN to reconstruct the banners, and then the target banners detection is executed again to get the new four corner points for replacing the tracked corners. If there is no overlap, the banners' tracking locations remain the same.

After the virtual logos occlusion checking, each tracked banner is also checked whether a soccer player blocks one of the four corners or not. If a corner is blocked by a player, the mean value of the other three corners' speeds is calculated to replace this blocked corner's speed. This fixes the soccer player blocking problem for the tracking.

The tracking process is done at this stage. The outcome is the four corners locations of all banners in the current video frame. Then, all tracked banners are checked to fix the four corner coordinates for banners on the video edge. Finally, the tracked banners are replaced by the virtual AD.

If a video frame is in a long-shot scene, the MODNet is triggered to segment the soccer players and extract their pixels. After the virtual AD pasting on the detected/tracked banners, the soccer players' pixels are pasted back by using the segmentation mask. This makes the occluded soccer players in front of the virtual AD.

After that, the virtual logos' pixels are extracted by using the virtual logo mask (manually labelled). These pixels are also pasted back onto the AD replacement frame. This displays the virtual logo on top of the virtual AD.

In the last stage, the detected/tracked target banners' information is collected for the next frame processing. If the next frame is a scene change, the target banners information is deleted. Otherwise, it is reminded for processing.

The processed new video frame (with AD replacement result) is output to create a new video. This is the entire process of the combination model.

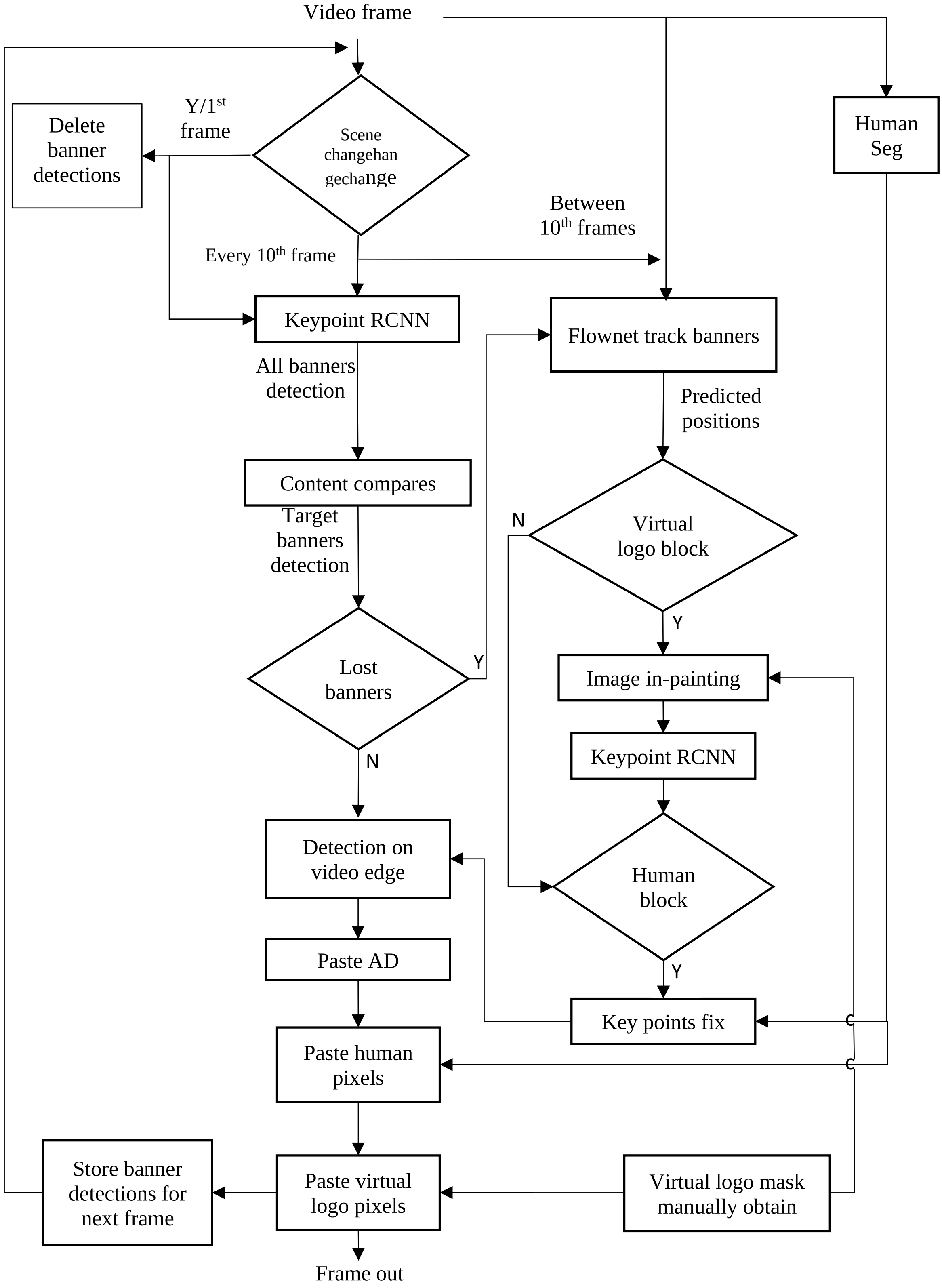


Fig. 11. The dynamic background AD replacement model process structure.

IV. Result and Discussion

The combination (AD replacement) model, which is introduced in Section III, includes four Neural Networks: Keypoint R-CNN-PF, Flownet v2, MODNet and AOT-GAN. To test the model, a full soccer game video (about 108 min) is used. Image data is collected from this *demo* video. Because different neural networks need different types of datasets, this section introduces data collection and neural network training. In addition, our proposed AD replacement model is applied to the entire video. After that, FOUR 5-mins videos are randomly selected from the demo video for evaluation. These 4 videos are used to measure the performance of the AD replacement model. Furthermore, we provide a short video to show an example of AD replacement. See the “Supplementary_Resouce1.mp4”.

A. *Training of Neural Networks*

Three of the four neural networks require data collection and training. They are Keypoint R-CNN-PF, AOT-GAN, and MODNet. The computer used for training is AMD Ryzen 7 2700 CPU, 16GB RAM and GPU of GTX 1070Ti. The programming language was Python with Pytorch. Flownet uses the pre-trained weights which are from the NVIDIA Flownet source code obtained from GitHub.

For Keypoint R-CNN-PF, 230 images are collected from different time periods of the demo video. 100 images have clear banners; another 100 images include banners blocked by virtual logos (team information, scores and TV logo). The other 30 images are for validation. 15 of them have clear banners; 15 include blocked banners. The blocked banner images (100 for training, 15 for validation) are input to AOT-GAN to reconstruct the blocked banners. After that, the re-draw images and the clear banner images are used to train Keypoint R-CNN-PF. All the banners on each image are labelled with four corners, including the reconstructed banners.

For AOT-GAN, 5 video clips are collected from the demo video. They are from different time periods. On average, each video clip has about 300 continuous frames (images). In these video clips, all banners are moving in the background. In total, around 1500 images are collected. These images are processed to prepare a dataset following Fig. 6. Each has broken banner content.

For MODNet, 180 images are collected from different time periods of the demo video. They all include banners which are occluded by soccer players. These 150 images are used for training. Another similar 30 images are also collected for validation. The 180 images are labelled for the soccer players; and then, they produce the football players’ masks which follow the instructions in Fig. 9.

After the training of the three neural networks, they are applied to the AD replacement model. The demo video is used for the final evaluation of the performance.

B. *Full Soccer Competition Video Application and Experiment Results*

After the training of the neural networks, the AD replacement model is applied to the entire 108-minute demo video. The target banners on this video are replaced by “Zyviz” AD. The target banner is randomly chosen. They are shown in Fig. 12. These banners are cropped from the demo video. They are also used as reference images for target banner detection.

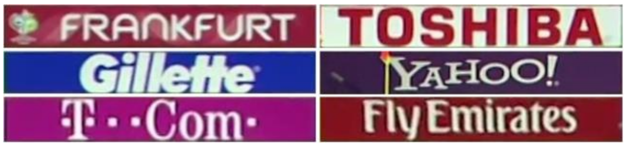
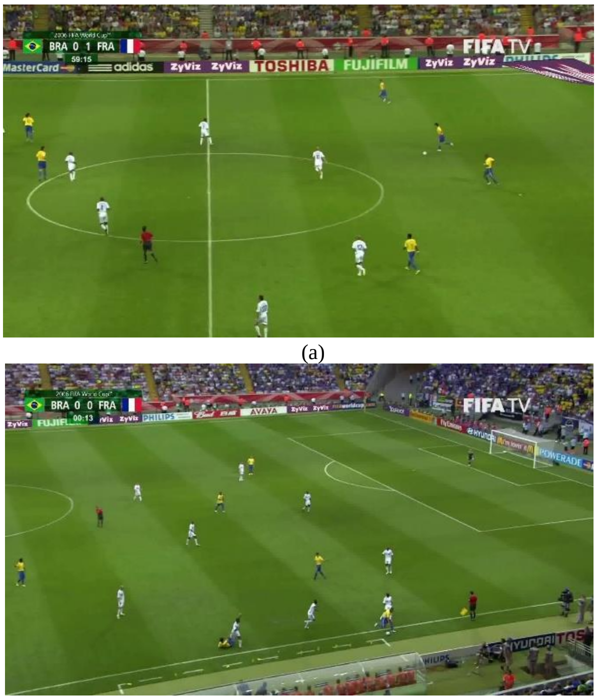


Fig. 12. The chosen target banners’ reference images. These are used for target banner detection.

The AD replacement model is programmed by using Python language. It processes the video frame by frame. The processing time is about 35 hours. The output video is the AD replacement. The target banners are replaced by “Zyviz” AD. The evaluation process is virtually observing each frame of the output video and recording the corrections and mistakes. However, the 108-minute video with 30 framerates has 194,400 frames. It takes a long time to measure the performance. Therefore, to measure the performance of our proposed model, 4 random 5-mins video clips are taken from the output video. The result is checked frame by frame.



(a)

(b)

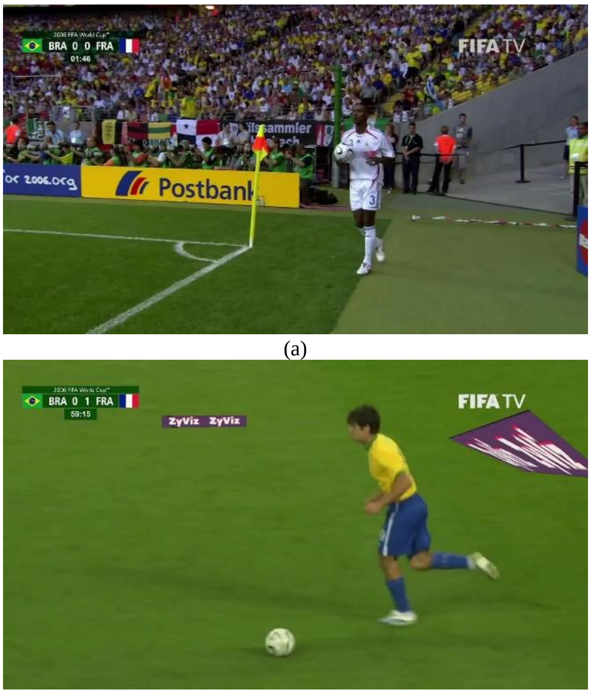
Fig. 13. The results on long-shot scene video frames. (a) 2 TPs, 1 FN and 1 FP (b) 3 TPs.

In the long-shot scene, the banners are displayed on the fence of the soccer field. It is the main camera view for AD replacement. In each frame, if a “Zyviz” AD replaces a target banner, it is a true positive (TP). If a “Zyviz” AD is pasted on a non-target banner, it is a False Positive (FP). If a target banner is not replaced by a “Zyviz” AD, it is a False Negative (FN). In Fig. 13(a), the banners of “Frankfurt” and “Gillette” are correctly replaced by the “Zyviz” AD, so there are 2 TPs. However, the “Toshiba” banner is not replaced. So, this is 1 FN. The banner of "Philips" is not the target banner but replaced. Therefore, this is 1 FP. Similarly, Fig. 13(b) has 3 TP with 3 target banners replaced correctly. The"Toshiba" banner on the left edge of the video frame is also counted as 1 TP.

The results of the long-short scene are shown in Table I. Three of the four video clips have a True Positive Ratio (TPR) greater than 80%. Only video clip 4 has 67.25% TPR. The main reason is the motion blur in these video clips. All 4 video clips have a Positive Predictive Value (PPV) greater than 96%. The best result is in video clip 2, which is 99.96%. The worst result is in clip 4, which is 96.58%.

TABLE I. THE LONG-SHOT SCENE RESULTS

Clips.	Frame No.	Scene No.	TP	FN	FP	TPR (%)	PPV (%)
1	5110	28	11,654	1501	233	88.59	98.03
2	4421	27	9268	1746	4	84.15	99.96
3	4284	19	8399	1417	144	85.56	98.31
4	5027	26	8550	4164	303	67.25	96.58
Total	18,842	100	37,871	8828	684	81.10	98.23


(a)

(b)

Fig. 14. The examples of short-shot scene results. (a) A frame result of 1 TN. (b) A frame result of 1 FP.

In the short-shot scene, there is no AD replacement. However, the "Zyviz" virtual AD may be mistakenly pasted on a frame of this scene. Thus, this is a False Positive (FP) frame. If a frame of a short-shot scene is not pasted by any "Zyviz" AD, it is a True Negative (TN) frame. Fig. 14(a) shows the 1 TN. There is no AD replacement on the frame. Fig. 14(b) shows 1 FP.

Table II indicates the results of the short-shot scene. All of the four video clips have a True Negative Ratio (TNR) greater than 87%. Particularly, video clips, 2 and 3, have TNR 95.4% and 97.5% respectively. The worst result is shown in clip 1, which is 87.5%.

It is difficult to measure the results numerically of soccer occlusion on the target banner AD replacement. The soccer player is very small, the evaluation IoU matrix can only compare the segmentation region pixels to the ground truth. Some small parts of the body, such as the head, legs and feet, which are not segmented well, still produce a high value of IoU. However, the virtual results show that this loss of small parts affects the audience's watch experience. For example, if a player's legs (legs are very thin in the video) are not segmented well, the AD replacement video shows the player kicking the ball with no legs, which is abnormal. Therefore, the scope of the segmentation evaluation method is not enough to measure the human occlusion performance on AD replacement.

TABLE II. THE SHORT-SHOT SCENE RESULTS

Clips	Frame No.	Scene No.	TN	FP	TNR(%)
1	3741	28	3273	468	87.5
2	4752	27	4533	219	95.4
3	4614	20	4499	115	97.5
4	4037	27	3566	471	88.3
Total	17,144	102	15,871	1273	92.57

Therefore, in this paper, only a virtual result is used to show the performance. A good example of the soccer occlusion on AD replacement is shown in Fig. 15. The football player segmentation has a high accuracy. Some heads, shoulders and arms can be displayed correctly in front of the "Zyviz" AD.

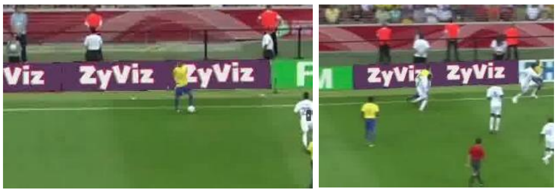

(a) (b)

Fig. 15. The example results of the football player occlusion on AD replacement.

C. Previous Works Comparison

The proposed AD replacement (combination) model was applied on the 108-minute test video continuously without pause and interruption. The related works, which did detection or segmentation on a static background, are not suitable for this dynamic background test video. There are no output results for comparing the proposed AD replacement model.

In addition, the live streaming AD replacement projects require camera intrinsic and extrinsic parameters. The 108 mins test video was taken in 2008. There are no camera parameters to use. Therefore, it is impossible to collect AD replacement results with those projects.

In summary, none of the previous works can be applied to compare to the proposed combination model for AD replacement performance.

D. Discussion

This research uses the soccer match video as an example to show the performance of AD replacement in a dynamic background. To apply the proposed model to other types of dynamic background sports videos, data collection is required. The data format is introduced in section IV.A. Three neural networks need to be trained, which are Keypoint R-CNN, AOT-GAN and MODNet. The Flownet v2 is used to get optical flow, it is not necessary to be trained. This paper does not show the implementation and results of other types of videos.

In the proposed AD replacement model, the four neural networks complete their own task individually. Keypoint R-CNN-PF is for target banner detection. Flownet v2 tracked the target banners' movement. The AOT-GAN reconstructed the banners which are blocked by the virtual logos. MODNet segments the soccer players to output masks. The AD replacement model utilizes the outputs of the neural networks to overcome their limitations. Keypoints R-CNN-PF cannot detect the target banners in every frame. Flownet v2's tracking helps to keep the detection of banners in every frame. Flownet v2 has a small error between continuous frames, it cannot run for a long time. Therefore, the tracking only runs 10 continue frames, then, Keypoint R-CNN-PF's detection is triggered in every 10th frame to fix the tracking error. Banners blocked by the virtual logos cannot be detected well, so AOT-GAN reconstructs the banners for detection. Although AOT-GAN cannot reconstruct banner content 100%, the target banners' edges are reconstructed well. This is enough for banners detection by Keypoint R-CNN-PF. Finally, AD replacement is affected a lot by human occlusion with banners. MODNet segments human pixels. These pixels can be pasted back to their original region after virtual AD pasting. In summary, the four neural networks cooperate very well in the proposed combination model. The total performance of the proposed model is 81.10% TPR, 98.23% PPV and 92.57% TNR. The total Successful Rate (SR) is calculated by using all of the TP, TN, FP and FN:

$$SR = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Based on the above equation, the total success rate is 83.33%. This shows the proposed model overcomes the limitations of the four neural network outputs and makes the AD replacement possible on the dynamic background banners on the video.

The proposed AD replacement model still has disadvantages. The Keypoint R-CNN-PF is not perfect for each banner's detection. When the output was abnormal (such as Fig. 13(a)), this would keep the abnormal AD replacement in 10 frames. Because the tracking runs in 10 frames.

Furthermore, the scene change detection was not always accurate. If an actual scene change was not detected, the proposed model would continue to track banners' regions, which causes the virtual AD incorrectly to keep pasting on the video. This mistake is shown in Fig. 14(b).

The soccer player occlusion solution in the AD replacement does not have a suitable evaluation method. The reason is explained in Section IV.B. It is difficult to prove the proposed model's performance.

E. Future Directions

One important future work is to improve scene change detection. The proposed AD replacement model relies on comparing the content of two continuous frames with HSV colour. It is better to use a deep learning method to improve accuracy. In the future, the proposed AD replacement model may combine a deep neural network for scene change detection with the other four networks.

Another possible improvement is to increase the accuracy of the four neural networks. All the neural networks, including Keypoint R-CNN-PF, Flownet v2, AOT-GAN and MODNet should be analyzed to find the possible modifications. If each neural network can be optimized and improved, the whole proposed model can perform better than before.

In addition, the soccer player occlusion solution requires a suitable evaluation model. The initial idea is to collect good occlusion results and broken occlusion results (losing head, arms, or legs). Then, these results are used to analyze which parts of losing affect the audience's watching experience.

Finally, the proposed AD replacement model utilizes the outputs of the four neural networks for the combination. It is possible to build a simple interface (or protocol) for new deep neural networks changing in the four tasks, including banner detection, tracking, in-painting and human segmentation. The interfaces make it easy to change new neural networks for each task. For example, in the future when the technology grows, a new in-painting neural network may appear which is better than AOT-GAN. If the proposed model has an interface for easily changing the new neural network to replace the AOT-GAN. The in-panting task can be better. Therefore, building interfaces for neural networks changing in the proposed AD replacement model can keep the model growing.

V. Conclusion

This paper proposes a dynamic background product replacement model that combines four different types of neural networks for AD replacement in social media videos. The proposed model applies virtual AD to multiple banners with dynamic backgrounds in the videos. It consists of four neural networks: object detection (Keypoint R-CNN-PF), optical flow measurement (Flownet V2), image in-painting (AOT-GAN), and image segmentation (MODNet). The findings demonstrate that in the long-shot scene, most video clips have TPR greater than 81%, and PPV greater than 96%. In the short-shot scene, most video clips have a TNR greater than 87%. Totally, the AD replacement on a dynamic background tends to be successful. The successful rate is 83.33%.

In the future, the research will follow the improvement directions, which are introduced in the last section, to increase the accuracy of the four tasks and the scene change detection. The proposed AD replacement model will keep growing. It will be easier to develop and get better for AD replacement on dynamic background videos.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

CY creates the methodology, completes the experiment and writes the original draft. FZ prepares and labels data. DZAH reviews the second version. PHJC edits the final version. PL contributes the conceptualization and supervises the progress. All authors had approved the final version.

References

[1] E. V. Karniouchina, C. Uslay, and G. Erenburg, "Do marketing media have life cycles? The case of product placement in movies," *Journal of Marketing*, vol. 75, no. 3, pp. 27–48, 2011.

[2] X.-H. Jiang, H.-L. Feng, and Y.-J. Dong, "Application of neural network in image detection of illegal billboards," in *Proc. International Academic Conference on Frontiers in Social Sciences and Management Innovation (IAFSM 2019)*, Atlantis Press, 2020, pp. 12–16.

[3] Á. Morera, Á. Sánchez, and A. B. Moreno *et al*., "SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities," *Sensors*, vol. 20, no. 16, p. 4587, 2020.

[4] L. Jin and G. Liu, "An approach on image processing of deep learning based on improved ssd," *Symmetry*, vol. 13, no. 3, p. 495, 2021.

[5] M. Sozzi, S. Cantalamessa, and A. Cogato *et al*., "Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms," *Agronomy*, vol. 12, no. 2, p. 319, 2022.

[6] F. Zheng, C. Yang, and P. H. J. Chong *et al*., "Deep learning algorithm for picture frame detection on social media videos," in *Proc. 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 2021, pp. 149–155.

[7] E. Ilg, N. Mayer, T. Saikia, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[8] Z. Ke, J. Sun, and K. Li *et al*., "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1140–1147.

[9] Y. Zeng, J. Fu, and H. Chao *et al*., "Aggregated contextual transformations for high-resolution image inpainting," *IEEE Transactions on Visualization and Computer Graphics,* 2022.

[10] A. Watve and S. Sural, "Soccer video processing for the detection of advertisement billboards," *Pattern Recognition Letters,* vol. 29, no. 7, pp. 994–1006, 2008.

[11] I. Iqbal, G. A. Odesanmi, and J. Wang *et al*., "Comparative investigation of learning algorithms for image classification with small dataset," *Applied Artificial Intelligence*, vol. 35, no. 10, pp. 697–716, 2021.

[12] K. He, G. Gkioxari, P. Dollár *et al*., "Mask R-CNN," in *Proc. the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[13] H. Zhang, Y. Gu, and S. Kamijo, "Orientation and occlusion aware multi-person pose estimation using multi-task deep learning network," in *Proc. 2019 IEEE International Conference on Consumer Electronics (ICCE)*, 2019, pp. 1–5.

[14] A. Westermann, P. Krayer, and A. Weiler. (2020). AdsISee: Advertisement detection and tracking for sponsorship evaluation in soccer matches. in *Proc. Workshop the EDBT/ICDT 2020 Joint Conference*, Copenhagen, Denmark. [Online]. vol. 2578: CEUR. Available: http://ceur-ws.org/Vol-2578/DARLIAP4.pdf

[15] S. Zhao, Y. Sheng, and Y. Dong *et al*., "Maskflownet: Asymmetric feature matching with learnable occlusion mask," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6278–6287.

[16] S. Jiang, D. Campbell, and Y. Lu *et al*., "Learning to estimate hidden motions with global motion aggregation," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9772–9781.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[18] G. Du, X. Cao, and J. Liang *et al*., "Medical image segmentation based on u-net: A review," *Journal of Imaging Science and Technology,* vol. 64, pp. 1–12, 2020.

[19] I. Iqbal, G. Shahzad, and N. Rafiq *et al*., "Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning," *IET Image Processing,* vol. 14, no. 10, pp. 1990–1998, 2020.

[20] Y. Li and H. Lu, "Natural image matting via guided contextual attention," in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11450–11457.

[21] H. Yu, N. Xu, and Z. Huang *et al*., "High-resolution deep image matting," in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3217–3224.

[22] S. Lin, A. Ryabtsev, and S. Sengupta *et al*., "Real-time high-resolution background matting," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8762–8771.

[23] J. Jam, C. Kendrick, and K. Walker *et al*., "A comprehensive review of past and present image inpainting methods," *Computer Vision and Image Understanding,* vol. 203, 103147, 2021.

[24] R. Liu, H. Deng, and Y. Huang *et al.*, "Fuseformer: Fusing fine-grained information in transformers for video inpainting," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14040–14049.

[25] Z. Li, C.-Z. Lu, and J. Qin *et al*., "Towards an end-to-end framework for flow-guided video inpainting," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17562–17571.

[26] R. Xu, X. Li, and B. Zhou *et al*., "Deep flow-guided video inpainting," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3723–3732.