Electronic Letters on Computer Vision and Image Analysis 20(1):117-136; 2021

Social Video Advertisement Replacement and its Evaluation in Convolutional Neural Networks

Cheng Yang[*,+], Xiang Yu[+], Arun Kumar[++], G. G. Md. Nawaz Ali[+++], Peter Han Joo Chong[*], Patrick Lam[+]

** Department of Electrical and Electronic Engineering, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, 34 St Paul St, Auckland, New Zealand*

+ Zyetric Technologies Limited, Room 112, WB, Building, Auckland University of Technology, Auckland, New Zealand

++ Department of Computer Science & Engineering, Odisha - National Institute of Technology, Odisha, India

+++ Department of Applied Computer Science, University of Charleston, 2300 MacCorkle Ave SE, Charleston, USA

Received 30th October 2020; accepted 14th May 2021

Abstract

This paper introduces a method to use deep convolutional neural networks (CNNs) to automatically replace advertisement (AD) photo on social (or self-media) videos and provides the suitable evaluation method to compare different CNNs. An AD photo can replace a picture inside a video. However, if a human being occludes the replaced picture in the original video, the newly pasted AD photo will block the human occluded part. The deep learning algorithm is implemented to segment the human being from the video. The segmented human pixels are then pasted back to the occluded area, so that the AD photo replacement becomes natural and perfect appearance in the video. This process requires the predicted occlusion edge to be closed to the ground truth occlusion edge, so that the AD photo can be occluded naturally. Therefore, this research introduces a curve fitting method to measure the predicted occlusion edge's error. By using this method, three CNN methods are applied and compared for the AD replacement. They are mask of regions convolutional neural network (Mask RCNN), recurrent network for video object segmentation (ROVS) and DeeplabV3. The experimental results show the comparative segmentation accuracy of the different models and DeeplabV3 shows the best performance.

Key Words: Deep Learning, Image Processing, Image Segmentation, Video Advertisement Replacement.

1 Introduction

Nowadays, Internet social (or self-media) videos are getting popularly. Many people upload their videos on social media platforms, such as on YouTube, Facebook and Twitter. The embedded advertisement (AD) [1, 2] has great business value based on this social phenomenon. The conventional AD embedded approach just uses video editing software (such as Photoshop) to paste the AD photos on the videos. This approach is easy to embed AD photo in front of objects on the video. However, if an AD needs to be embedded behind an object, such as human being, then a video editing tool needs a significantly huge amount of effort to edit

Correspondence to: royang@aut.ac.nz, robert.yang@zyetric.com, ych2tj@gmail.com

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1347
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

the video due to the occlusion problem. This research work enables the AD photo to occlude with the human being automatically on a video.

The main technique of this research is the segmentation of the human being from the occluded part in the video. The human being detection and segmentation is a popular topic in computer vision (CV). In the early years, human detection focuses on finding features of human being on images. Methods are created for the detection of features. These include background subtraction [3], colour detection [4] and texture detection [5]. Over the past years, many machine learning (ML) methods are used to detect human being. These methods include Adaboost [6], random forest [7] and support vector machine (SVM) [8]. The ML algorithms choose and combine multi-methods for human being detection. The latest CV algorithm for the human being segmentation is convolutional neural networks (CNNs) [9, 10]. The CNNs find features by using convolutional layers. This algorithm is smarter and more accurate for human being detection and segmentation.

In addition, the AD photo occlusion requires high accuracy of human segmentation, so that the AD photo can be occluded naturally. Notably, the predicted occlusion edge needs to be closer to the ground truth occlusion edge. In this paper, the intersection over union (IoU) [11] is used to evaluate the deep learning performance. However, this method focuses on the area of overlap between prediction and ground truth. It cannot measure the occlusion edge prediction.

There are two main contributions in this research. The first contribution is to solve the AD replacement occlusion problem using deep CNN. The AD photo is pasted on a video frame, which also covers the occlusion part of a human being. Then, a CNN model is used to segment human being pixels from the similar original video frame. After that, segmented human being pixels are pasted back to the video frame, in which AD photo has already pasted. Three popular CNN methods are implemented: (i) mask region-based convolutional neural network (Mask RCNN), (ii) recurrent net for video object segmentation (ROVS) and (iii) Deeplab Version 3 (DeepLabV3). The detail of these techniques is explained in Section 3.

The second contribution of this paper is to introduce a novel evaluation method, based on the curve fitting algorithm, to compare the three CNN methods. To evaluate the performance of CNN methods on the occlusion solution for the AD photo replacement, this research uses IoU [11] to measure segmentation region. However, the IoU focuses on the accuracy of the human being area segmentation. This research work requires not only the area segmentation accuracy but also the occlusion edge (or human shape on the occlusion part) accuracy. This paper introduces a special curve fitting method to measure the occlusion edge prediction errors.

The rest of the paper is organized as follows. Section 2 overviews the related work on AD photo replacement and human being segmentation. Section 3 proposes the AD photo replacement method to avoid occlusion using three different CNN techniques. Section 4 introduces our proposed evaluation method to evaluate occluded edge prediction performances. Section 5 presents the experimental results and discussion. Also, this section presents a comparison among 3 different deep learning models. Finally, Section 6 concludes the paper with future work directions.

2 Related works

This paper focuses on an AD photo replacement method to overcomes the human being occlusion problem in videos. This requires the human segmentation performed at high accuracy. Particularly, the human body shape needs to be segmented efficiently and accurately. Otherwise, the occlusion between the AD photo and human being will not be natural. To the best of our knowledge, this area has not been explored enough and we are one of the few to explore this research area.

In the very early years, many computer vision techniques were proposed for human detection and segmentation, such as background subtraction [3, 12, 13], colour detection [14-16], texture detection [5, 17, 18]. The background subtraction is to separate the static background and moving foreground. The moving human being is segmented from the static background using Mixture of Gaussian (MOG) [19], frames difference [12], optical flow [20] and Bayesian formulation [21]. The background subtraction segments pixels which belong to the moving objects on a video. This method is widely used in the indoor CCTV system, because a human being is a moving object on a CCTV video. However, for outdoor CCTV and

social media videos, there are many non-human moving objects. These objects can be mis-detected as human being. Furthermore, the background subtraction is sensitive to the object's moving speed. The model should be set up with some parameters for adapting the detected object's speed. If the parameters used are not suitable, there will be so-called "ghost" pixels in the detection result. Therefore, the segmentation accuracy is not robust enough.

The colour detection uses skin colour to detection human being. Methods including Gaussian mixture model [22], Fusion of colour [4] and colour classifiers [23] are all based on the skin colour. The disadvantage of colour detection is that the clothes have different colours and patterns, which reduce the detection accuracy significantly. Also, texture detection uses different objects textures to distinguish human from other objects. It is suitable for finding the contour of human beings. The most common method is the dynamic texture [24, 25]. However, the texture feature is difficult for people wearing different pattern clothes. Finally, different methods are also combined to improve detection and segmentation accuracy. The first example is the combination of the colour and texture. This method segments different objects, including human being [26]. Texture attributes and colour features are modelled with a multivariate finite mixture model [27]. This model can be used to segment different objects. Another example is to apply background subtraction on colour pixels and texture features [28, 29]. A multi-layer background subtraction model is used to take moving objects colour and texture features [29]. However, the limitation of these combined methods includes the complex background, noise and shadows which lead to a very non-accurate human segmentation.

Over the past years, machine learning techniques have become the mainstream to apply for human being detection, such as Adaboost cascade classifiers [30], random forest [31] and support vector machine (SVM) [8]. The ML uses different features, such as colour, texture, contour and pattern, to detect human being. Different methods analyse different features in different approaches. Adaboost cascade classifiers [6, 32-34] uses multiple classifiers (building from features) for human being detection. However, the detection accuracy is not very well in the image with complicated colour, pattern and features. The random forest uses features to distinguish the human and other objects in binary operation [7], such as clothing segmentation [35], multi-limb human segmentation [36] and occluded human detection [37]. Each feature is used in one layer for the binary operation [38]. Therefore, the whole detection model is a binary tree. However, the segmentation accuracy is not very good, if the features between human and other objects are not obvious. The SVM analyses different features as vectors [8, 39, 40]. The features constitute a multi-dimensional space [41]. The SVM uses the space to assemble the object's features optimally. This is the application of the SVM for objects detection, including human being [42]. In summary, ML algorithms automatically operate features of human beings to find the best performance. However, the ML algorithms should be provided with a large amount of training dataset and extracted the suitable features for the detection model. If features are not enough for the detection model, the accuracy will be low. Conversely, providing enough features to achieve high accuracy is difficult.

Nowadays, deep learning algorithms are becoming popular for human being segmentation. Especially, the CNN becomes the most successful methods for segmentation [43]. In the CNN models, the feature extracted by using multiple convolution layers with large example data. Therefore, CNN does not need to be provided with auditable features. Recently, fully convolutional neural network (FCNNs) based methods [44, 45] show effective feature generating. They become the most popular choice for segmentation. After that, different types of CNNs are created for segmentation. Some CNNs are utilized for image segmentation. The most popular CNNs are Mask RCNN [46] and DeepLab [47-49]. Mask RCNN is based on the region-based convolutional neural network [50-52] in which the neural network model is generated by adding the parallel branch on the top of RCNN. It provides segmentation by generating binary mask for each class object. The DeepLab is another successful model in the field of image segmentation. This model improves the segmentation of the object by extracting multiple scales with the help of atrous spatial pyramid pooling (ASPP). Some CNNs have good performance on video base segmentation. The one-shot video segmentation [53, 54] is an example which uses CNN architecture to tackle a semi-supervised video segmentation. In this method, the input is the segmentation of the first frame and the output is the masks of the object in the next frames of the sequence. Another example is to use recurrent neural network (RNN) [55, 56] which combines with CNN for video segmentation. This is clearly to be shown in [55] which is called RVOS. In this method, CNN is used to extract features from each frame of video. The RNN is utilized to remember the features

along the video frame sequence for the next video frame segmentation. This combination approach learns the spatial-temporal dependence of a video. Therefore, it can do the zero-shot video segmentation.

In this research, the CNN algorithm is utilized to segment human being, which occludes the pictures. The algorithm can automatically segment a human being with the highest accuracy compared to other algorithms. After that, the AD photo can replace the picture in the video to eliminate the occlusion problem. This approach requires occluded human shape to be segmented efficiently and accurately. To evaluate the occlusion edge (or shape) prediction, the normal evaluation methods, such as Intersection of Union (IoU) [11] and contour accuracy [26, 57] may not be suitable. Since IoU focuses on the accuracy of the human being region segmentation, it cannot show the accuracy of the shape prediction. The contour accuracy uses receiver operating characteristic (ROC) to analyze number of predicted shape pixels on the true edges of the image. Normally, this method is utilized to evaluate image edge detection, but it may not be suitable for this research. The CNN segments human shape that may not be exactly on the true edge, but a little bit far from the true edge. Therefore, the evaluation method should show how far the predicted shape from the image true edge.

3 The proposed occlusion solution

In this research, the CNNs are utilized to segment human being, which occludes the pictures, in the videos. The algorithm can automatically segment a human being with the highest accuracy compared to other algorithms. After that, the AD photo can replace the picture in the video without the occlusion problem.

3.1 Method for AD Photo Replacement with Occlusion.

(a) (b)

Figure 1: An example of AD photo replacement: (a) The original video frame image (*O*). (b) Two AD photos replace the pictures from the video frame (*P*). This replacement is completed manually.

The occlusion problem of the AD photo replacement is shown in Figure 1. The original video frame image in Figure 1 (a) shows that the girl on the left occludes the picture behind her. If the two AD photos are pasted on the video frame manually as shown in Figure 1 (b). Then, the AD photo in the left blocks the human body, which overlays the girl's head. Our proposed solution is to edit the video frame by frame and automatically bring the blocked part of a human head back to the video. To solve the occlusion problem illustrated in Figure 1, we propose to combine both the image processing method and the CNN method. Our proposed technique has three steps as shown in Figure 2.

The first step is to paste the AD photos to the picture area in the original video frame in Figure 1 (a). Then, the original video frame now becomes a new video frame which has the AD pasted images (called image *P*) as shown in Figure 1 (b).

The second step is to segment the human being from the original video frame. The CNN is utilized to do this. The detail description of CNNs used in this paper is explained in section 3.2. The result is a human mask image (called image *M*) which masks the individual human pixels with a value of "1", but the background pixels have the value of "0".

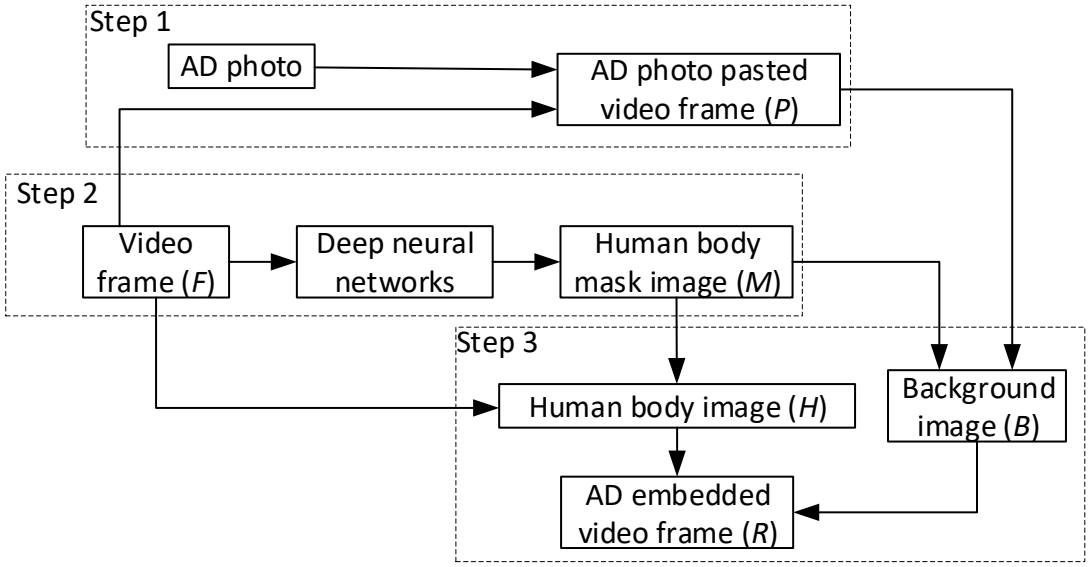


Figure 2: The proposed AD logo replacement framework

The third step is to collect the human body pixels (called image (*H)*) using the human mask image, image (*M*), from the original video frame. After that, these human body pixels, image (*H*), are pasted back to the AD pasted image (called image (*R*)). This makes the AD photo to occlude with the human being.

In sum, the human mask image, image (*M*), is used to get the human body pixels, image (*H*), from the original video frame. After that, the human mask image, image (*M*), is also used to get background pixels, (called image (*P*)), from the pasted video image. Finally, the human body, image (*H*), and the background, image *B,* are merged to get the resulting image, image (*R*). Figure 3 shows the resulting image (*R*), which embeds the two AD photos to a picture frame; and the AD photo on the left occludes with the human body.

Figure 3: The final result image (*R*). This is used for the final video frame

Figure 3 shows the final resulting image. The human segmentation is obtained using DeeplabV3, which gives very high accuracy. However, some other CNN models produce the results which may not be as good as using DeeplabV3. If the model does not have high accuracy for human segmentation, the shape of the human being will be displayed abnormally in the occlusion area as shown in Figure 9 in Section 4. This research chooses three CNN methods for comparison to solve the AD occlusion problem which is explained in the next section.

3.2 Application of convolutional neural networks

The occluded part of the human being should be accurately segmented so that the AD logo replacement can be observed naturally. In this research, three CNNs, Mask R-CNN, RVOS and DeeplabV3, are employed

to solve the AD occlusion problem. Their segmentation results are compared to find which network is the most suitable for the AD replacement.

3.2.1 Mask R-CNN

In Mask R-CNN, the segmentation is done after object detection [46]. This CNN detects objects based on Fast R-CNN. The core network in the Mask R-CNN is the "restnet101" which includes 100 convolutional layers. After that, it produces objects segmentation with 28×28 mask to represent pixels of the segmented objects. In this research, the Mask R-CNN model is pre-trained with COCO dataset [58]. However, in social media videos, the picture area for AD replacement mostly occluded with human beings. Therefore, the human being images are collected to retrain the Mask R-CNN. Human being images are mainly collected from COCO dataset. The total number of images is 6000, which is split into training (5000 images) and validation (1000 images) dataset. After the training, the network is implemented for the human being's segmentation.

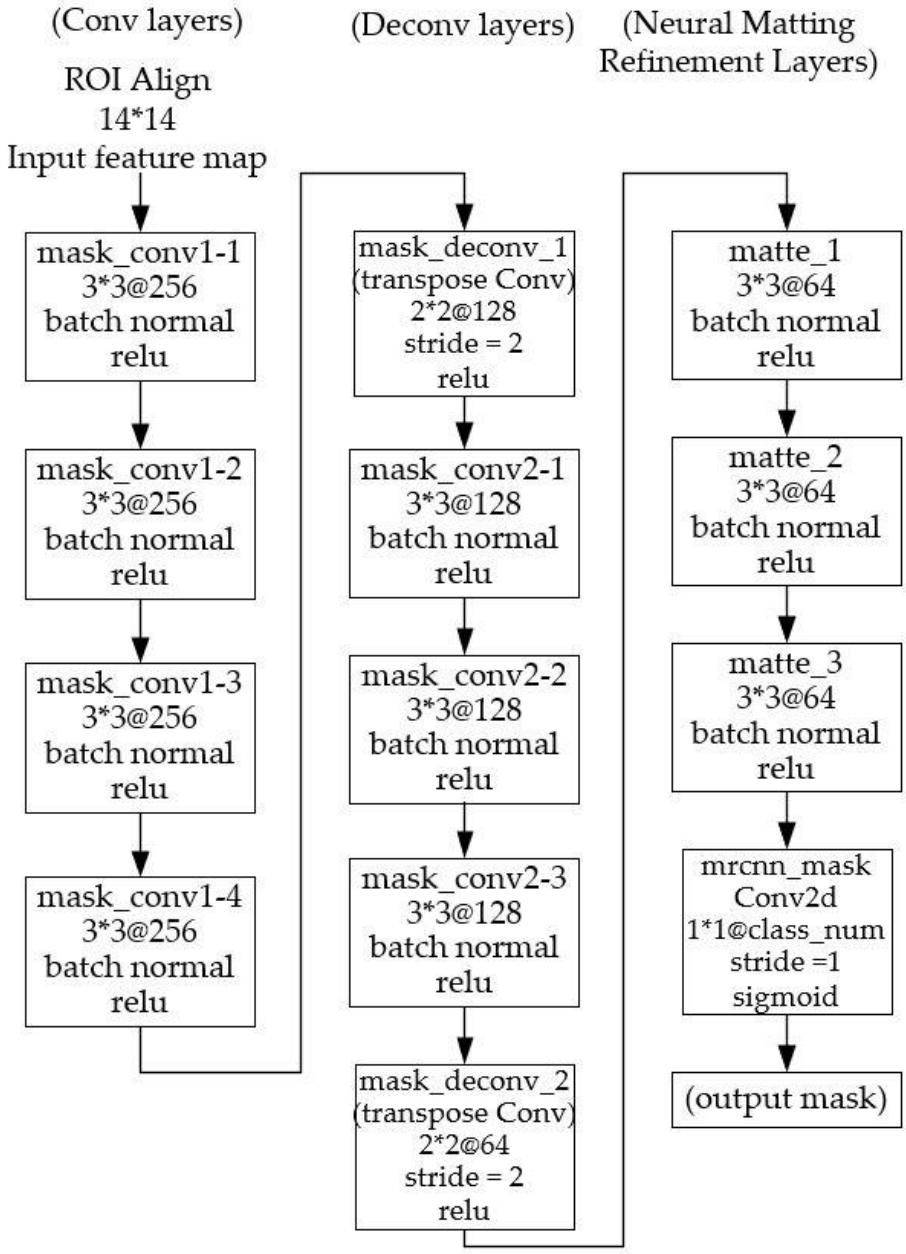


Figure 4: The new mask branch layers which replace the original mask layers of Mask R-CNN

The Mask R-CNN has a mask brunch which includes convolutional layers and a deconvolutional layer. This brunch produces masks of objects. However, it produces 28×28 mask output which is too small to satisfy the AD logo replacement because a social media video has 1920×1080 resolution. The output mask for the video shows the segmentation edge is not smooth enough. Therefore, mask branch in Mask R-CNN is modified to produce a 56×56 mask by adding another deconvolutional layer. Finally, mask output can segment human beings with smooth edge. The modified network is shown in Figure 4. The Mask R-CNN is applied by using Tensorflow in Linux system.

3.2.2 RVOS

In RVOS (recurrent network for video object segmentation) method [55], the convolutional neural network (CNN) and recurrent neural network (RNN) are combined to do the segmentation from a video frames flow. The CNN extracts features and the RNN uses video sequences to "remember" the features flow. The structure of the network looks like an U-Net. The RVOS has encoder-decoder architecture. The encoder architecture consists of a ResNet-101 [59] model. The weights of this model are pre-trained by ImageNet [60] dataset. It does instance segmentation by predicting a sequence of masks. The input x_t of the encoder is an RGB image, which corresponds to frame t in the video sequence, and the output $f_t = \{f_{t,1}, f_{t,2}, \ldots, f_{t,j}\}$ is a set of features at different resolutions. The decoder is designed as a hierarchical recurrent architecture of ConvLSTMs [61] which leverages the different resolution of the input features $f_t = \{f_{t,1}, f_{t,2}, \ldots, f_{t,j}\}$, where $f_{t,j}$ are the features extracted at the level j of the encoder for the frame t of the video sequence. The output of the decoder is a set of object segmentation predictions $\{P_{t,1}, , \ldots, P_{t,i}, \ldots, P_{t,N}\}$, where $P_{t,i}$ is the segmentation mask of the object i at frame t. The paper introduces a spatial and temporal recurrence. The output $l_{t,i,j}$ of the j-th ConvLSTM layer for object i at frame t depends on the following variables: (1) the features of f_t obtained from the encoder from frame t; (2) the preceding j-1-th ConvLSTM layer; (3) the hidden state representation from the previous object i-1 at the same frame t ($l_{t,i-1,j}$), which will be referred to as the spatial hidden state; (4) the hidden state representation from the same object i at the previous frame t-1 ($l_{t-1,i,j}$), which will be referred to as the temporal hidden state; (5) the object segmentation prediction mask $P_{t-1,I}$ of the object i at the previous frame t-1. The formulas are shown below:

$$l_{input} = [B_2(l_{t,i,j-1})|f'_{t,j}|P_{t-1,i}] \tag{1}$$

$$l_{state} = [l_{t,i-1,j}|l_{t-1,i,j}] \tag{2}$$

$$l_{t,i,j} = ConvLSTM_j(l_{input}, l_{state}) \tag{3}$$

where B_2 is the bilinear up-sampling operator by a factor of 2 and $f'_{t,j}$ is the result of projecting $f_{t,K}$ to have lower dimensionality via a convolutional layer. Equation (3) is applied in a chain for $j \in \{1, \ldots, m_b\}$, being m_b as the number of convolutional blocks in the encoder. $l_{t,i,0}$ is obtained by considering:

$$l_{input} = [f'_{t,0}|P_{t-1,i}] \tag{4}$$

For the first object, l_{state} is obtained as below:

$$l_{state} = [Z|l_{t-1,i,j}] \tag{5}$$

where Z is a zero matrix to represent that there is no previous spatial hidden state for this object. The RVOS officially provides two pre-trained models: one is trained by DAVIS 2017; the other is trained by YouTube-VOS.

These models show good performance for human segmentation, but other objects are segmented as noise. The result in social media video may segment other objects on the background with human beings. These objects sometimes are in the picture frames which are not desired to be segmented. The problem is shown in Figure 5 (a). Since the original video's frame may have the copyright, it only shows the problem parts.

(a) (b)

Figure 5: The RVOS problem and solution. (a) The problem is that the false positive is painting picture in the background. (b) The solution result after segmentation.

In Figure 5 (a), the red part belongs to the human body. The black ellipse beside the human body is a painting part which belongs to the background of the picture. This false positive is segmented as part of the human body. The result is from an RVOS model which is trained by DAVIS2017 dataset [62]. In this dataset, the mask does not include only human bodies, but also other objects along with human bodies, such as backpacks, skateboards and bikes.

To solve this problem which is shown in Figure 5 (a), we modify the DAVIS2017 dataset to only include the mask of human beings. All other objects are not masked. This new dataset is used to train the RVOS model. One of the image results is shown in Figure 5 (b). It can be seen that the false positive part of Figure 5 (a) is removed a lot. The RVOS is applied with Pytorch in Linux system.

3.2.3 DeepLabV3

The DeepLabV3 neural network is proposed by [49]. The special feature is the Atrous Separable Convolution (ASC) layers which can control the resolution of features and adjust filter's field-of-view in order to capture multi-scale information and generalize standard convolution operation. In the case of 2-dimensional signals, for each location n on the output feature map y and a convolution filter c, atrous convolution is applied over the input feature map x as follows:

$$y[n] = \sum_j x[n + r \cdot j]c[j] \tag{6}$$

where the atrous rate, r, determines the stride with the sampled input signal. The standard convolution is a special case of rate $r = 1$. The filter's field-of-view is adaptively modified by changing the rate value. This is also called dilated convolution or Hole Algorithm.

The ASC significantly reduces the computation complexity of the proposed model while maintaining similar (or better) performance. With Atrous Spatial Pyramid Pooling (ASPP), the output features from ASC layers are encoded to multi-scale contextual information. After a proposed decoder method in [49], the features are recovered to output a predicted mask.

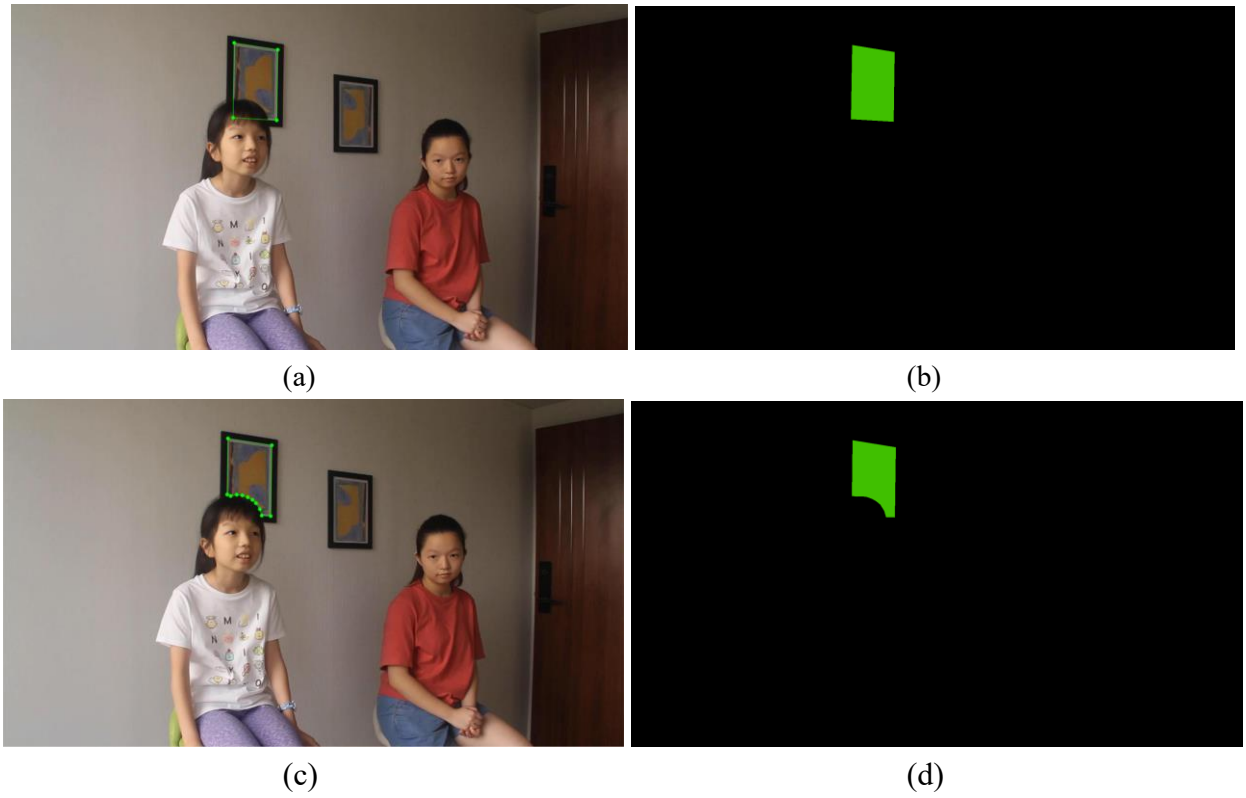
In this research, the Deeplabv3 network is implemented by using GluonCV. The model is pre-trained by COCO dataset. The experimental results show that this network segments human being from social media video better than the other 2 CNNs.

The above three CNN models are implemented in our AD photo replacement model. These three models' performances are compared with the proposed evaluation method described in Section 4.

4 Curve fitting evaluation method for occlusion

Ten social media videos are collected for the test. Each video includes 50 continuous frames. All these videos have the areas of picture frame which are suitable to be replaced by AD photo. However, these replacement areas are occluded with human heads/bodies. To enable the evaluation, the ground truth mask of

the replacement area and the occluded area need to be labelled manually. Totally, 1000 video frames are labelled for the evaluation. Figure 6 shows an example of the labels for one of the videos.



(a) (b)

(c) (d)

Figure 6: The test data labels. (a) The whole picture frame's label. (b) The corresponding mask of (a). (c) The replaced picture area label. (d) The corresponding mask of (c).

In Figure 6, there are two types of labels. Figure 6 (a) is the label of the entire picture frame. This label includes the human occluded area. This label is only done once on each video, because the picture in the video is a static object. Figure 6 (c) displays the label which only mask the picture area. The human being occluded area is not in the label. This is done in all frames of the 10 videos. The mask of the entire picture frame and the mask of the picture area is shown in Figure 6 (b) and Figure 6 (d), respectively. These two masks are also used to produce the human occluded area, which is indicated in Figure 7.

Figure 7: The human occluded area from the labelled masks

One of the evaluation methods in this research is the Intersection of Union (IoU) There are two IoU values are used in this experimental test; one is the AD replacement area segmentation IoU between the ground truth mask and the prediction mask areas (ADR IOU), and the other one is the human occlusion area segmentation IoU between the ground truth mask and the prediction mask areas (HMO IoU).

Figure 8 illustrates these two concepts of IoU. In Figure 8 (a), the green solid lines show the AD replacement ground truth area. The red dashed lines show the segmentation area of AD replacement. This is

the Mask RCNN result. The ADR IOU is calculated from these two areas. In Figure 8 (b), the green solid lines are the ground truth human occluded area and the red dashed lines indicate the segmented human occluded area which is the result of Mask R-CNN. The HMO IoU is calculated from these two areas.

(a) (b)

Figure 8: The two IoU values for evaluation. The ground truth area is green; and the segmentation area is red. (a) The AD replacement area IoU (ADR IoU). (b) The human occluded area IoU (HMO IoU).

The ADR IOU shows how well the AD logo is replaced. The higher the value is close to 100%, the better the prediction model performs. However, if the true occlusion area is very small, the ADR IOU can be very high; even the occlusion part is not segmented correctly because the intersection area occupies most of the union area. Therefore, the HMO IoU is also applied. If the true occlusion area is small, but the predicted occlusion area is none, then the HMO IoU value is zero. This is much clearer to show the prediction model's performance because the intersection area can be very small.

Since each video has 50 frames, there will be 50 ADR IoUs and 50 HMO IoUs. Each video's ADR/HMO IoU consists of the mean and standard deviation. For both the ADR IoU and HMO IoU, a higher mean value resembles more accurate prediction. Thus, AD logo replacement becomes more natural. A high value in the standard deviation shows the prediction of human occlusion part has variation through the frame flow. Thus, the segmentation area jumps bigger and smaller on the resulting video.

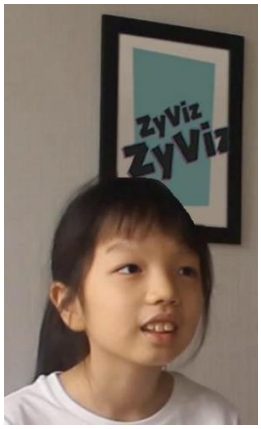
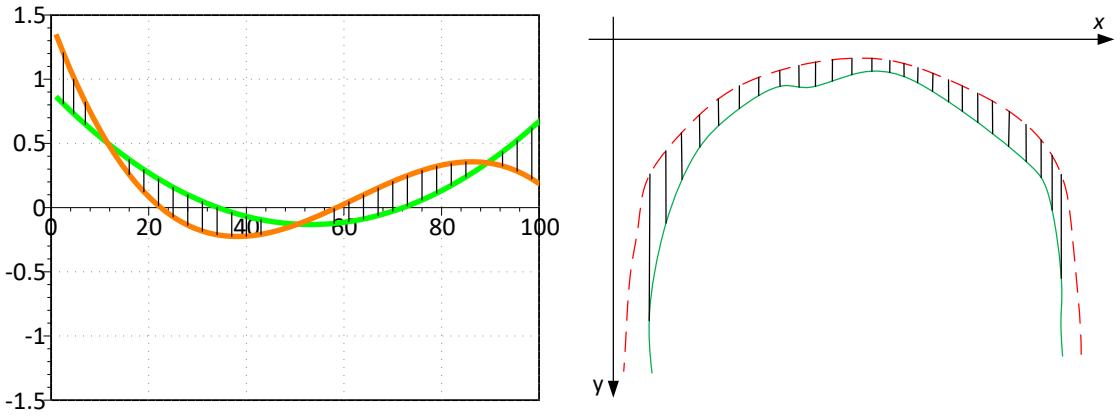


Figure 9: The AD photo replacement result. The result does not look natural, because the human head's shape is not segmented accurately

In AD replacement, if the edge between human occlusion area and picture frame area is not detected very accurate, AD replacement does not look natural as shown in Figure 9. Therefore, the prediction mask edge is required to close to the ground truth mask edge. To investigate this performance, a curve fitting method is introduced.

In statistics, the curving fitting is to calculate errors between two curves with one coordination. For example, there are two curves on an x-y axis (Figure 10 (a)). The method takes sample points from two curves with the same x-axis. Then, it calculates errors between y-axis. After that, the RMS of the errors is

calculated to show the curve fitting performance. Figure 10 (a) displays the errors between the curves on the diagram. The short black vertical lines represent the errors between samples points on the two curves. These errors are collected to calculate the RMS which evaluates how much the two curves fitting to each other. If the two curves are very different from each other, the RMS value will be very high. Otherwise, if the two curves are very similar, the RMS value will be small.



(a) (b)

Figure 10: The statistic curve fitting. (a) The curve fitting method applied on the two curves. (b) The curve fitting method applies on comparing the edges of prediction and ground truth.

However, the curve fitting used in this research is different from the traditional statistics curve fitting technique [63]. An edge of the human head which occludes the AD photo is not a curve going along the x-axis. It may be rounded as a semicircle. Thus, the error calculation can be significantly large. Figure 10 (b) shows an example of a human head edge curve. The green curve is the ground truth edge; and the red dashed line indicates the prediction mask edge. If the errors are gained based on the approach from Figure 10 (a), which are the black short lines between the ground truth and prediction edges, the errors on the left and right sides of these edge curves are very big. Thus, the RMS has a very high value. However, the edge curve from the prediction is closed to the ground truth edge as seen in Figure 10 (b). This should not produce a very high RMS value.

To avoid this problem, this research uses a new method to define the errors between ground truth edge and prediction edge. The labelled ground truth edge is used as a standard curve. All pixels (coordination) on this edge are taken as sample points. Each pixel is used to find the shortest distance (in pixel unit) pixel which is from the curve of the predicted edge. This shortest distance is a sample error between the ground truth edge and predicted edge. After finding the errors between the ground truth edge and predicted edge, the RMS values are calculated from these errors. If the ground truth edge is shorter than the predicted edge, some of the pixels on the predicted edge may not be used. Thus, they have been left. Conversely, if the ground truth edge is longer than the prediction edge. Some different pixels on the ground truth curve may find the similar shortest distance pixel from the prediction curve. However, the values of these distances are mostly different. Thus, this may not become a big problem.

An example is shown in Figure 11 (a). The two edge curves are similar to the curves in Figure 10 (b). In Figure 11 (a), the short black lines show the shortest distances (errors) for some sample points between the ground truth edge and the predicted edge. These errors are used to calculate the RMS. This RMS value can be used for evaluation of prediction of human edge. If the prediction edge is close to the ground truth edge, the RMS value is smaller. If the prediction edge is very different from the ground truth edge, the RMS value is greater.

Figure 11 (b) shows another example of calculating errors between the ground truth edge (green curve) and the prediction edge (red curve). The black lines indicate the shortest distances (errors) of some sample points. These errors are used to calculate the RMS value. Figure 11 just gives two examples to explain the new curve fitting approach. In the test experiment, all pixels in the ground truth edge are chosen as sample points for calculating the errors and the corresponding RMS.

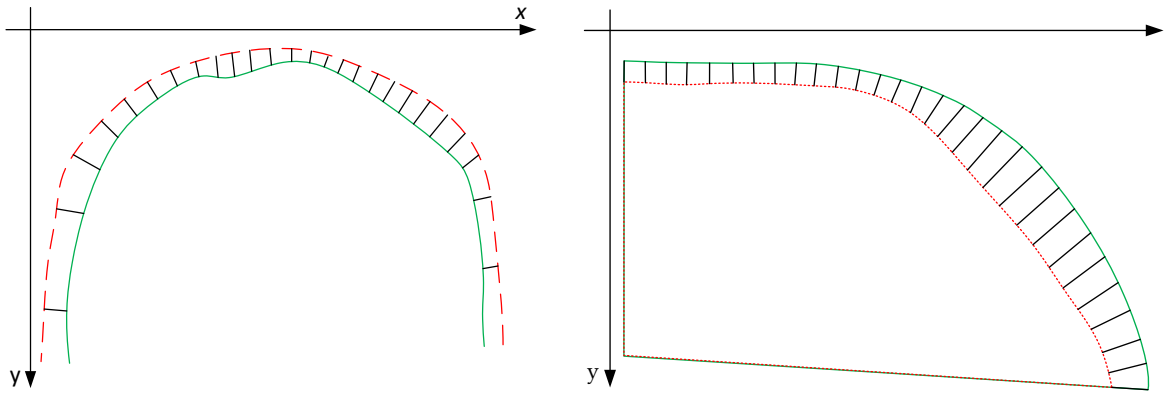


Figure 11: Using the pixels shortest distances, which are the errors, for evaluating the human edge prediction. The black lines represent the kind of errors. However, they are just used to show examples about what are the shortest distances looks like, they are not the exact shortest distances. (a) The example of errors with the method from this research, which improve the curve fitting problem in Figure 10 (b). (b) A curve fitting example from a frame of a test video. The curves are from Figure 8 (b). The errors do not include the background picture's edge. The picture's edge is manually annotated, errors must be zeros.

The errors' RMS between ground truth edge and the predicted edge is applied in each video frame. Each video has 50 frames. Therefore, each video can get 50 RMS values. These values are collected to calculate the mean and standard deviation. For each video, the lower mean RMS value shows that the prediction is more accurate. The standard deviation may show the stability of the prediction process in the 50 frames. A high standard deviation value means the prediction edges are fluctuation through the frame flow. However, the lower standard deviation may be produced by the prediction edges up or down to the ground truth edge with same errors among different frames. Conversely, this is not common among the test videos prediction results.

20 videos are used to test three CNN methods (Mask R-CNN, RVOS and DeepLabV3) for the occlusion. Each CNN method produces edges of segmentation prediction on the 20 videos. The prediction edges are used to calculate the RMS's mean and standard deviation with the ground truth edges. Therefore, there are 10 RMS's mean and standard deviation values for each CNN method. The three CNN methods evaluation is shown in Section 5.

5 Experimental results and discussion

The entire model from Section 3.1 is created by Python 3.6 on Ubuntu Linux system. The hardware parameters of the PC are AMD Ryzen 7 2700 3.20GHz, 16GB RAM and NVIDIA Geforce GTX 1070 Ti. The Mask R-CNN is applied in the environment of Tensorflow. The RVOS is built with PyTorch. The DeepLabV3 is implemented on GluonCV package. The test results are shown in three tables.

Table 1 shows the Mask R-CNN results. The column of "Pixels No." shows the number of the average pixels of ground truth human edge which is occluded in the background by replaced picture. The pixels do not include the background picture's boundary. An example is indicated in Figure 11 (b). These pixels on the edge curves are the sample points in each frame. The RMS is the root mean square of errors between the prediction and ground truth edge curves for each frame in each video. The table shows the mean (Mean) and standard deviation (Stdv) of the RMS of each video. The columns of "ADR IoU" and "HOM IoU" also show the mean value (Mean) and standard deviation (Stdv) of IoUs.

In Table 1, the Mask RCNN has the RMS of errors around 10 pixels in 8 videos (videos of 2, 3, 5, 6, 10, 13, 14 and 16). Particularly, videos 2 and 5 have the Stdvs of RMS greater than 4.9. These high values show the high fluctuation of the mask areas' prediction. The AD photo replacements will not be natural in these videos. Videos 1, 4, 7, 9, 11, 15, 18, 19 and 20 have RMS of errors around 4.0 ~ 7.5 pixels. These

videos also perform the non-natural AD replacement, but they are not as obvious as the above 8 videos. Other 3 videos (No. 8, 12 and 17 have RMS around 3.0 ~ 4.0, the AD replacement is about natural, but not perfect. In the results of ADR IoU, most of the videos have values higher than 0.9 (90%), only two videos (No. 2 and 4) have values less than 0.9. In the results of HMO IoU, most of the videos have greater than 0.7 values. However, videos 4 and 6 have values only 0.261 and 0.388, respectively. This is because the human occluded area is very small in these two videos. Thus, the Mask R-CNN only segments a very small area; even it does not segment any occluded area sometimes.

Videos	Frames	Pixels No.	Mask RCNN					
			RMS		ADR IoU		HMO IoU	
			Mean	Stdv	Mean	Stdv	Mean	Stdv
1	50	551	4.92	2.8138	0.991	0.0060	0.926	0.0433
2	50	253	9.73	4.9246	0.806	0.0563	0.839	0.0249
3	50	238	10.61	1.2084	0.951	0.0082	0.749	0.0242
4	50	145	6.81	2.3180	0.970	0.0149	0.261	0.2921
5	50	205	11.36	6.6479	0.946	0.0345	0.770	0.1178
6	50	173	12.40	1.9771	0.987	0.0019	0.388	0.0968
7	50	188	6.28	0.7009	0.971	0.0030	0.773	0.0253
8	50	128	3.76	1.2587	0.993	0.0025	0.840	0.0612
9	50	222	6.47	1.2929	0.931	0.0182	0.840	0.0320
10	50	513	13.60	3.4836	0.982	0.0054	0.822	0.0630
11	50	404	7.42	3.432	0.922	0.0067	0.728	0.0454
12	50	172	3.12	1.428	0.953	0.0083	0.933	0.0461
13	50	306	10.42	4.5368	0.962	0.0317	0.816	0.0907
14	50	250	8.97	3.9225	0.817	0.052	0.843	0.0332
15	50	197	5.22	1.543	0.984	0.0121	0.678	0.1934
16	50	254	9.63	1.1044	0.958	0.0078	0.763	0.0343
17	50	330	3.57	1.4772	0.988	0.0127	0.937	0.0415
18	50	645	4.37	2.736	0.952	0.0065	0.938	0.0247
19	50	235	5.48	0.8938	0.942	0.0174	0.886	0.0347
20	50	943	4.92	0.7456	0.955	0.0142	0.911	0.0371

Table 1: Mask RCNN results

Table 2 shows the ROVS results for AD photo replacement. The No. of video frames and No. of edge pixels are similar to Table 1, these two columns are not shown in Table 2. In RVOS, 14 videos have RMS values less than 3.0. In these videos, the AD replacement is natural and perfect. However, all RMS mean values in Mask RCNN are greater than 3. Therefore, RMS of errors in RVOS results is better than Mask RCNN. Four videos (9, 11, 14 and 18) has RMS mean value around 4.0. In these four videos, the contrast between the background picture and the human being occlusion part is not obvious, so that the RVOS does not perform perfectly. 19 videos have fewer RMS standard deviation values than Mask RCNN. This shows the RVOS predicts the occlusion edges with less fluctuation through the video frame sequence. However, video 10 has a significant poor result in which the mean value is 14.73 and standard deviation of 7.8715. This is because, in video 10, the background picture includes black colour, which is similar to human hair. Thus, RVOS cannot segment the human area accurately. In the ADR IoU results, all videos have mean values greater than 0.95. However, Mask RCNN has 6 videos less than 0.95. In the HMO IoU results, most of the videos have mean values around 0.9. Particularly, videos 1, 2, 3, 12, 15, 16, 17 18 and 19 (9 videos in total) have the values greater than 0.95. However, videos 4 and 10 have lower values of 0.779 and 0.819,

respectively. In this case, RVOS has better results than Mask RCNN. In summary, Table 2 shows that the RVOS segments the occlusion part and perform the AD replacement more accurately than Mask RCNN.

Videos	RVOS					
	RMS		ADR IoU		HMO IoU	
	Mean	Stdv	Mean	Stdv	Mean	Stdv
1	2.08	0.3953	0.996	0.0015	0.963	0.0133
2	2.86	1.1136	0.958	0.0159	0.965	0.0108
3	1.70	0.3222	0.990	0.0022	0.951	0.0104
4	1.93	0.4546	0.990	0.0028	0.779	0.0996
5	2.19	0.2094	0.990	0.0012	0.941	0.0058
6	1.96	0.9734	0.997	0.0012	0.896	0.0386
7	2.48	0.1979	0.987	0.0009	0.903	0.0065
8	2.85	0.7956	0.995	0.0010	0.878	0.0256
9	3.97	1.4850	0.958	0.0120	0.894	0.0368
10	14.73	7.8715	0.982	0.0083	0.819	0.1002
11	3.74	1.0765	0.981	0.0043	0.911	0.0186
12	1.08	0.1592	0.991	0.0031	0.967	0.0106
13	2.75	0.5864	0.990	0.0037	0.937	0.0070
14	5.36	2.5392	0.957	0.0200	0.930	0.0330
15	1.31	0.3553	0.996	0.0007	0.952	0.0173
16	2.09	0.4154	0.988	0.0026	0.951	0.0092
17	2.26	0.3542	0.975	0.0048	0.959	0.0109
18	4.10	1.0359	0.984	0.0035	0.968	0.0065
19	2.11	0.1801	0.962	0.0035	0.956	0.0031
20	3.06	0.3379	0.979	0.0056	0.936	0.0172

Table 2: RVOS results

Table 3 indicates the DeepLabV3 results of AD photo replacement. In the RMS of errors, 18 videos have values less than (or close to) 3.0, which is more than RVOS. This shows the DeepLabV3 performs segmentation better than RVOS. Only, videos 8 and 10 have values of 4.27 and 4.06, respectively. Video 8 includes the human hair which is transparent to the background picture, so DeepLabV3 doesn't process it well. Although the background is black (similar to the human hair) in video 10, the performance of DeepLabV3 is much better than RVOS and Mask RCNN. In the standard deviation of RMS, DeepLabV3 has 19 videos' values smaller than Mask RCNN and has 11 videos' values smaller than RVOS. This indicates that the DeepLabV3 predicts the occlusion edges with less fluctuation than both Mask RCNN and RVOS. In the ADR IoU results, 19 videos have mean values greater than 0.95. Particularly, 17 videos have these values greater than 0.98, which RVOS has 14 videos. Only video 2 has ADR IoU mean value less than 0.95, but the value is 0.941 which is close to 0.95. In the HMO IoU results, 14 videos have mean values higher than 0.95, which is more than RVOS (9 videos). Video 4 has the lowest value of 0.764, which is slightly lower than the result of RVOS. Because the occlusion part in this video is very small, the HMO IoU is sensitive to the non-segmented occlusion part. However, both DeepLabV3 and RVOS perform much higher values than Mask RCNN, which has only 0.261. Video 10's HMO IoU value is 0.953 which is better than the results from RVOS and Mask RCNN. In summary, Table 3 indicates that the DeepLabV3 has the best performance among the three convolutional neural network models.

Videos	DeepLabV3					
	RMS		ADR IoU		HMO IoU	
	Mean	Stdv	Mean	Stdv	Mean	Stdv
1	1.55	0.3149	0.997	0.0021	0.972	0.0181
2	2.81	2.3665	0.941	0.0305	0.950	0.0290
3	1.73	0.5472	0.990	0.0027	0.954	0.0139
4	2.76	1.5916	0.984	0.0130	0.764	0.0644
5	0.72	0.1717	0.996	0.0011	0.979	0.0061
6	1.50	0.2412	0.998	0.0002	0.916	0.0107
7	1.10	0.1895	0.993	0.0006	0.949	0.0047
8	4.27	2.1017	0.993	0.0032	0.847	0.0664
9	1.82	0.3837	0.982	0.0030	0.951	0.0107
10	4.06	2.3479	0.994	0.0026	0.953	0.0248
11	2.17	0.4456	0.988	0.0015	0.942	0.0073
12	0.84	0.2386	0.993	0.0030	0.974	0.0135
13	2.07	1.2642	0.994	0.0011	0.957	0.0191
14	1.87	0.5582	0.985	0.0041	0.978	0.0058
15	1.20	0.3118	0.996	0.0009	0.952	0.0166
16	2.27	0.4836	0.988	0.0028	0.952	0.0088
17	2.16	0.4572	0.977	0.0037	0.962	0.0125
18	1.79	0.1811	0.993	0.0007	0.987	0.0014
19	2.04	0.2005	0.962	0.0048	0.956	0.0039
20	3.19	0.2272	0.975	0.0071	0.929	0.0113

Table 3: The DeepLabV3 results

In the above three tables, the results of the 20 videos are shown. The Mask RCNN has a lower mean value of HMO IoUs and higher mean values of RMS errors in most of the videos. The performance of Mask RCNN is the worst. The RVOS has RMS mean values significantly lower than Mask RCNN in most of the videos. Furthermore, the mean values of ADR IoU and HMO IoU are higher than Mask RCNN for most of the time. However, RVOS has a high value of RMS in one video. The DeeplabV3 model has low values of RMSs and high values of ADR IoUs in most videos comparing to the other two models. In HMO IoU results, this model has the highest values; except that one video is slightly lower than RVOS. Thus, the DeeplabV3 model has the best performance. This model is the main technology to be implemented for AD photo replacement, so far, for the Zyetric company.

The curve fitting method uses RMS of the pixels difference to measure whether the AD replacement can be virtually natural or not. It focuses on the edge of the segmentation. It is more efficient than the IoU evaluation parameters. Because the IoU method focuses on the segmentation area of the occlusion part, but the edge of the occlusion part more relates to how natural the AD replacement. For example, video 10 has high values of ADR IoU on all three deep neural network models (Mask RCNN: 0.982; RVOS: 0.982; DeepLabV3: 0.994). However, the AD replacement results on Mask RCNN and RVOS are very bad. In this case, the RMS values on these two models are very high (13.60 on Mask RCNN and 14.73 on RVOS). The RMS values clearly show the bad performance of these two models. Another example is on video 4 HMO IoU. The occlusion area on this video is small, which effects the values are low on RVOS and DeepLabV3 models (0.779 and 0.764 respectively). However, the AD replacement is virtually natural on this video with these two models. In this situation, the curve fitting RMS values on these two models are less than 3.0 (RVOS: 1.93; DeepLabV3: 2.76), which clearly indicate the successful results of the AD replacement. In summary, the curve fitting method efficiently measures the AD replacement performance for CNN models.

Figure 12 shows the vision results with the three CNN methods. The example is from the first frame of video 3. In Figure 12, the Mask RCNN has the worst result. However, the results of RVOS and DeepLabV3 are similar. This figure just shows an example of an occlusion solution. It cannot be used to compare the CNN methods. The three tables above can numerically indicate which CNN method is the best.

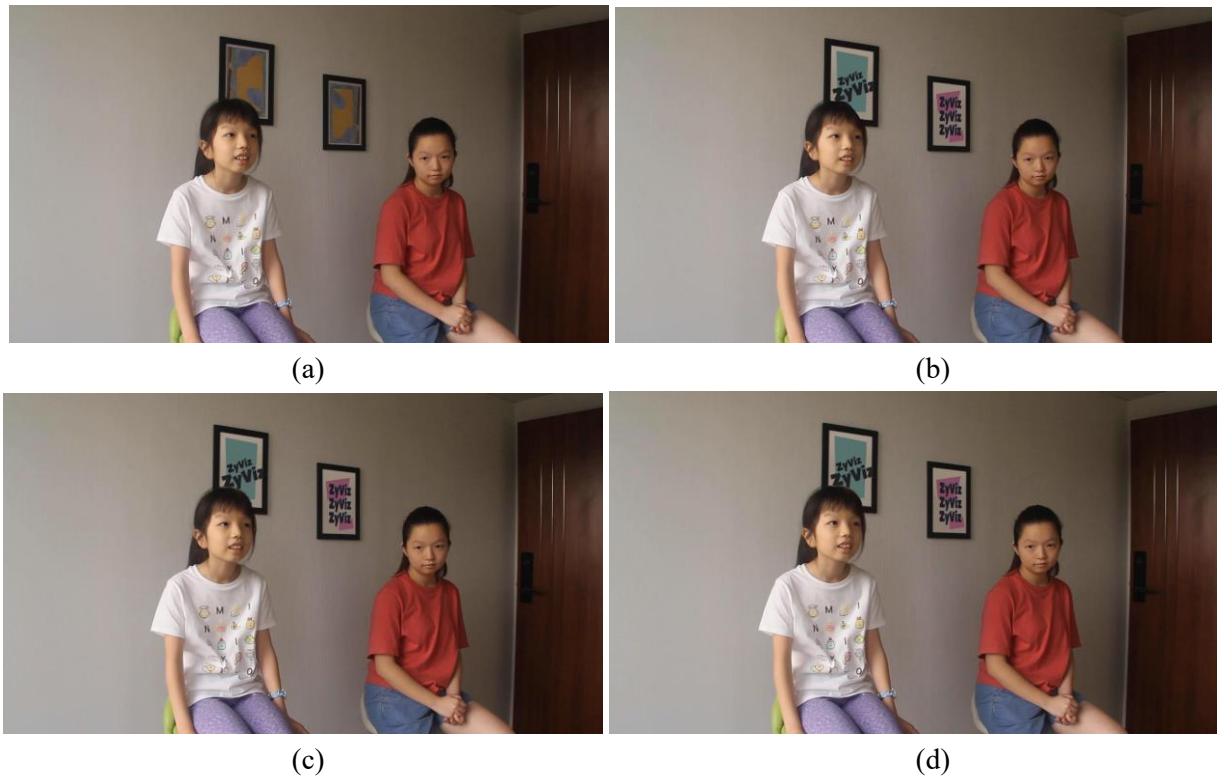


Figure 12: Vision results of the AD replacement. (a) The original image. (b) The result of implementing Mask RCNN. (c) The result of implementing RVOS. (d) The result of implementing DeepLabV3.

6 Conclusions

This paper introduces a novel application of using CNN methods to solve the occlusion problem for advertainment (AD) photo replacement (or AD replacement model) on social media videos. In a social media video, if a human being occludes a picture which is required to be replaced by an AD, this advert photo will block the occluded human body. This research has used CNN methods to segment human being area, and then paste this area back to the video. This approach successfully solves the occlusion problem.

In this paper, three deep learning models have been implemented for AD photo replacement. The results are compared by using IoUs and a special curve fitting method. The IoUs evaluates the region of the segmentation. The special curve fitting method is proposed to evaluate the segmentation shape. This indicates whether the final AD photo is embedded naturally or not. According to the test videos, the curve fitting method is more efficient to measure the performance of the AD photo replacement.

The evaluation results indicate that the proposed AD replacement model successfully implement CNN methods to solve the occlusion problem. Particularly, the AD replacement model with DeepLabV3 performs the best results on 20 test videos. The AD replacement model with RVOS is the second-best approach. However, the AD replacement with Mask R-CNN is not able to replace AD logo naturally on most of the tested social media videos.

Even though the AD photo is replaced with a picture from a social media video naturally, the replaced area of a picture is given to the model manually. This task is labor extensive if the camera moves during the video recording. Our future work is to detect the potential replaced area by using CNN methods as well.

In addition, our proposed occlusion solution still needs more video data to verify and confirm the efficiency and robustness. For example, videos including different objects occlusion should be collected, such as animals occluded with each other or cars occluded with posters. Also, different methods for different object segmentation can be used to test our occlusion solution. Furthermore, our proposed special curve fitting evaluation method needs to be compared with other evaluation methods. Even although the contour accuracy [26, 57] is not suitable for this research of evaluating the occlusion avoidance solution, it still needs more experiments to prove our proposed evaluation method. For example, we can use the contour accuracy to evaluate the performance of other occlusion avoidance technique based on CNN. If the performance evaluation of contour accuracy is not accurate, then this can prove our proposed evaluation method to be better because our proposed evaluation method focuses on the distance between the predicted occlusion edge and the ground truth occlusion edge.

References

[1] H. Zhang, X. Cao, J. K. L. Ho, and T. W. S. Chow, "Object-Level Video Advertising: An Optimization Framework," *IEEE Transactions on Industrial Informatics,* 13(2):520-531, 2017, doi: https://doi.org/10.1109/TII.2016.2605629.

[2] C. Luo, Y. Peng, T. Zhu, and L. Li, "An optimization framework of video advertising: using deep learning algorithm based on global image information," *Cluster Computing,* 22(4):8939-8951, 2019, doi: https://doi.org/10.1007/s10586-018-2024-3.

[3] R. S. Rakibe and B. D. Patil, "Background subtraction algorithm based human motion detection," *International Journal of scientific and research publications,* 3(5):2250-3153, 2013.

[4] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognition,* 40(6):1771-1784, 2007, doi: https://doi.org/10.1016/j.patcog.2006.11.010.

[5] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human activity recognition using a dynamic texture based method," in *BMVC*, 1:2, 2008, doi: http://dx.doi.org/10.5244/C.22.88

[6] C.-C. R. Wang and J.-J. J. Lien, "AdaBoost learning for human detection based on histograms of oriented gradients," in *Asian Conference on Computer Vision*, Springer, 1:885-895, 2007, doi: https://doi.org/10.1007/978-3-540-76386-4_84.

[7] D. Tang, Y. Liu, and T.-K. Kim, "Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates," in *BMVC*, Citeseer, 1:1-11, 2012.

[8] A. H. Ahmed, K. Kpalma, and A. O. Guedi, "Human Detection Using HOG-SVM, Mixture of Gaussian and Background Contours Subtraction," in *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 4-7:334-338, 2017, doi: 10.1109/SITIS.2017.62.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* 521(7553):436-444, 2015, doi: 10.1038/nature14539.

[10] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks,* 61:85-117, 2015, doi: https://doi.org/10.1016/j.neunet.2014.09.003.

[11] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1:658-666, 2019.

[12] S. Saravanakumar, A. Vadivel, and C. G. S. Ahmed, "Multiple human object tracking using background subtraction and shadow removal techniques," in *2010 International Conference on Signal and Image Processing*, 1:79-84, 2010, doi: 10.1109/ICSIP.2010.5697446.

[13] Z. Jianpeng and H. Jack, "Real Time Robust Human Detection and Tracking System," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 1:149-149, 2005, doi: 10.1109/CVPR.2005.517.

[14] F. Gasparini and R. Schettini, "Skin segmentation using multiple thresholding," in *Internet Imaging VII*, 6061: International Society for Optics and Photonics, p. 60610F, 2006, doi: https://doi.org/10.1117/12.647446.

[15] M. Störring, H. J. Andersen, and E. Granum, "Skin colour detection under changing lighting conditions," in *7th Symposium on Intelligent Robotics systems*, Citeseer, 1999, doi: 10.1.1.28.7702

[16] D. Petrişor, C. Foşalău, M. Avila, and F. Măriuţ, "Algorithm for face and eye detection using colour segmentation and invariant features," in *2011 34th International Conference on Telecommunications and Signal Processing (TSP)*, 1:564-569, 2011, doi: 10.1109/TSP.2011.6043666.

[17] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using texture and local shape analysis," *IET biometrics,* 1(1):3-10, 2012, doi: https://doi.org/10.1049/iet-bmt.2011.0009.

[18] A. Conci, E. Nunes, J. J. Pantrigo, and Á. Sánchez, "Comparing Color and Texture-Based Algorithms for Human Skin Detection," in *ICEIS (5)*, 1:166-173, 2008.

[19] L. Li and J. Xu, "Moving human detection algorithm based on Gaussian mixture model," in *Proceedings of the 29th Chinese Control Conference*, 1:2853-2856, 2010.

[20] A. Fernández-Caballero, J. C. Castillo, J. Martínez-Cantos, and R. Martínez-Tomás, "Optical flow or image subtraction in human detection from infrared camera on mobile robot," *Robotics and Autonomous Systems,* 58(12):1273-1281, 2010, doi: https://doi.org/10.1016/j.robot.2010.06.002.

[21] E. How-Lung, W. Junxian, A. H. Kam, and Y. Wei-Yun, "A Bayesian framework for robust human detection and occlusion handling human shape model," in *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004.* 2:257-260, 2004, doi: 10.1109/ICPR.2004.1334150.

[22] R. Hassanpour, A. Shahbahrami, and S. Wong, "Adaptive Gaussian mixture model for skin color segmentation," *World Academy of Science, Engineering and Technology,* 41:1-6, 2008.

[23] S. Duffner and J.-M. Odobez, "Leveraging colour segmentation for upper-body detection," *Pattern Recognition,* 47(6):2222-2230, 2014, doi: https://doi.org/10.1016/j.patcog.2013.12.014.

[24] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Recognition of human actions using texture descriptors," *Machine Vision and Applications,* 22(5):767-780, 2011, doi: 10.1007/s00138-009-0233-8.

[25] T. d. Freitas Pereira *et al.*, "Face liveness detection using dynamic texture," *EURASIP Journal on Image and Video Processing,* 2014(1)2, 2014, doi: 10.1186/1687-5281-2014-2.

[26] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26(5):530-549, 2004, doi: 10.1109/TPAMI.2004.1273918.

[27] J.-S. Kim and K.-S. Hong, "Color–texture segmentation using unsupervised graph cuts," *Pattern Recognition,* 42(5):735-750, 2009, doi: https://doi.org/10.1016/j.patcog.2008.09.031.

[28] M. Harville, "A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models," in *European Conference on Computer Vision*, Springer, 1: 543-560, 2002, doi: https://doi.org/10.1007/3-540-47977-5_36.

[29] J. Yao and J.-M. Odobez, "Multi-layer background subtraction based on color and texture," in *2007 IEEE conference on computer vision and pattern recognition*, IEEE, 1:1-8, 2007, doi: https://doi.org/10.1109/CVPR.2007.383497.

[30] S. Ma and T. Du, "Improved Adaboost Face Detection," in *2010 International Conference on Measuring Technology and Mechatronics Automation*, 2:434-437, 2010 doi: 10.1109/ICMTMA.2010.184.

[31] A. Joshi, C. Monnier, M. Betke, and S. Sclaroff, "A random forest approach to segmenting and classifying gestures," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1-7, 2015, doi: 10.1109/FG.2015.7163126.

[32] S. Bakheet and A. Al-Hamadi, "A hybrid cascade approach for human skin segmentation," *Journal of Advances in Mathematics and Computer Science,* 1:1-14, 2016, doi: https://doi.org/10.9734/BJMCS/2016/26412.

[33] J. Xu, Q. Wu, J. Zhang, and Z. Tang, "Fast and Accurate Human Detection Using a Cascade of Boosted MS-LBP Features," *IEEE Signal Processing Letters,* 19(10):676-679, 2012, doi: 10.1109/LSP.2012.2210870.

[34] G. Wei, A. Haizhou, and L. Shihong, "Adaptive Contour Features in oriented granular space for human detection and segmentation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1:1786-1793, 2009 doi: 10.1109/CVPR.2009.5206762.

[35] N. Wang and H. Ai, "Who Blocks Who: Simultaneous clothing segmentation for grouping images," in *2011 International Conference on Computer Vision*, 1:1535-1542, 2011, doi: 10.1109/ICCV.2011.6126412.

[36] A. Hernandez-Vela *et al.*, "Graph cuts optimization for multi-limb human segmentation in depth maps," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1:726-732, 2012, doi: 10.1109/CVPR.2012.6247742.

[37] B. C. Ko, J. E. Son, and J.-Y. Nam, "View-invariant, partially occluded human detection in still images using part bases and random forest," *Optical Engineering,* 54(5):053113, 2015, doi: https://doi.org/10.1117/1.OE.54.5.053113.

[38] A. Joshi, C. Monnier, M. Betke, and S. Sclaroff, "Comparing random forest approaches to segmenting and classifying gestures," *Image and Vision Computing,* 58:86-95, 2017, doi: https://doi.org/10.1016/j.imavis.2016.06.001.

[39] Z. Lin and L. S. Davis, "A Pose-Invariant Descriptor for Human Detection and Segmentation," *Springer Berlin Heidelberg, in Computer Vision*, ECCV, Berlin 1:423-436, 2008, doi: https://doi.org/10.1007/978-3-540-88693-8_31

[40] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *IEEE 12th International Conference on Computer Vision*, 1:24-31, 2009, doi: 10.1109/ICCV.2009.5459205.

[41] L. Spinello and R. Siegwart, "Human detection using multimodal and multidimensional features," in *2008 IEEE International Conference on Robotics and Automation*, 1:3264-3269, 2008, doi: 10.1109/ROBOT.2008.4543708.

[42] L. Bertelli, T. Yu, D. Vu, and B. Gokturk, "Kernelized structural SVM learning for supervised object segmentation," in *CVPR*, 1:2153-2160, 2011 doi: 10.1109/CVPR.2011.5995597.

[43] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1:3193-3202, 2017.

[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1:3431-3440, 2015.

[45] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 39(12):2481-2495, 2017, doi: 10.1109/TPAMI.2016.2644615.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 1:2961-2969, 2017.

[47] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 40(4):834-848, 2018, doi: 10.1109/TPAMI.2017.2699184.

[48] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587,* 2017.

[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 1:801-818, 2018.

[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1:580-587, 2014.

[51] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 1:1440-1448, 2015.

[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 1:91-99, 2015.

[53] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1:221-230, 2017.

[54] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation," *IEEE Transactions on Cybernetics,* 50(9):3855-3865, 2020, doi: 10.1109/TCYB.2020.2992433.

[55] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i-Nieto, "Rvos: End-to-end recurrent network for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1:5277-5286, 2019.

[56] Z. Qiu, T. Yao, and T. Mei, "Learning Deep Spatio-Temporal Dependence for Semantic Video Segmentation," *IEEE Transactions on Multimedia,* 20(4):939-949, 2018, doi: 10.1109/TMM.2017.2759504.

[57] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1:724-732, 2016.

[58] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," Cham, 2014: *Springer International Publishing,* in Computer Vision, ECCV, 1:740-755, 2014, doi: https://doi.org/10.1007/978-3-319-10602-1_48.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1:770-778, 2016.

[60] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision,* 115(3):211-252, 2015, doi: https://doi.org/10.1007/s11263-015-0816-y.

[61] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 1:802-810, 2015.

[62] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675,* 2017.

[63] P. G. Guest and P. G. Guest, "Numerical methods of curve fitting," *Cambridge University Press*, 2012.