An Occlusion Handling Evaluation Criterion for Deep Learning Object Segmentation

C Yang1,*a, P H J Chong1,b and P Lam2,c

1 Department of Electrical and Electronic Engineering, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, 34 St Paul St, Auckland, New Zealand
2 Zyetric Technology Limited, Unit 513, 5/F, Building 19W, No. 19 Science Park West Avenue, Hong Kong, China

*royang@aut.ac.nz; b peter.chong@aut.ac.nz; c patrick.lam@zyetric.com

Abstract. This paper introduces a novel evaluation criterion to occlusion handling for deep learning object segmentation. The occlusion is defined as objects blocking each other on an image. It affects deep learning object segmentation. More and more researches focus on occlusion handling for object segmentation. However, these researches do not clearly show the evaluation of their occlusion handling method, because there is no suitable evaluation criterion. Traditionally, people just use images results or use the entire object boundary accuracy to show the occlusion handling of their methods. Conversely, these ideas cannot give a numerical evaluation focusing on occlusion handling. This research reports an evaluation criterion to measure occlusion handling performance. This evaluation criterion uses the shortest distances between the pixels from the ground truth occlusion edges and the segmentation (result) shape. The shortest distances are segmentation errors. Then, the average value of these errors is the final parameter for occlusion handling evaluation criterion. The experiment uses a deep learning based segmentation model as an example. It shows that this criterion (or method) successfully measures the occlusion handling for deep learning based object segmentation.

1. Introduction

Object segmentation is a popular topic in the computer vision field. In object segmentation, if objects block each other on an image, there is an occlusion problem. Figure 1 (a) shows an example, which the adult man is blocked by the child on his left side. The occlusion on images can cause a reduction of the segmentation accuracy. Occlusion is a difficult problem in the segmentation, because it leads to an overlap of object boundaries [1]. Many kinds of research try to handle the occlusion problem for object segmentation. However, these researches cannot clearly show how well the methods handle the occlusion. Because of this, it is not easy to see whether a method can efficiently handle the occlusion problem or not. Traditionally, occlusion handling researches just use some image results or the entire object boundary accuracy to show the occlusion handling of their methods.

Figure 1. An example of ground truth occlusion edge. (a) The original image. (b) The occluded objects masks. (c) The occlusion edge curve in binary image.
Gouda et al. [2] handled the occlusion problem in multiple human tracking on video processing. They improved Mean Shift Tracking algorithm to track the occlusion target. In their research, occlusion was detected by calculating the centre of mass of both objects and when the distance between them was zero. However, they did not show their experiment results. Fuzhen et al. [1] reported a method using selective shape priors for image segmentation under occlusion. They modified Nitzberg-Mumford-Shiota variational formulation to analyse prior knowledge of the shape of objects for finding occluded boundaries. In the experiment, they just used images to show their results. Xiong et al [3] derived an occlusion map for foreground objects and used the occlusion map as the “seeding” interactive input to an interactive image segmentation approach. They also only show some images results, but didn’t have any numeric evaluation approach to test their method. A Multi-Scale and Occlusion Aware Network was introduced in [4] for Unmanned Aerial Vehicles (UAV) based vehicle segmentation. In this research, a Multi-Scale Feature Adaptive Fusion Network could adaptively aggregate hierarchical feature maps from multiple levels. Then, a Regional Attention-based Triple Head Network was utilized to suppress background noise caused by occlusions. Although they used segmentation area IoU to show their segmentation results, their test did not focus on measuring the accuracy of occlusion handling. Chen et al [5] reported their method for overcoming the occlusion, which was called Graph Cut. Their method incorporated top-down category specific reasoning and shape prediction through convolutional neural networks (CNN) into an intuitive energy minimization framework. In their experiment, they took images which include occluded objects to test their method with segmentation IoU. Conversely, the objects also had non-occluded regions included in the overall evaluation. In summary, researches above did not clearly show the performance of occlusion handling. There is no legacy method focus on evaluation of occlusion handling on deep learning object segmentation.

The contribution of this research is to create an evaluation criterion to occlusion handling for the deep learning based object segmentation models. It uses the ground truth occlusion boundary (or edge) as the standard reference to test the performance of occlusion part segmentation. The ground truth occlusion boundary is shown in Figure 1 (c), which can be gained from the ground truth mask in Figure 1 (b). A deep learning based segmentation model predicts the segmentation of the objects on an image. If one object occludes to another one, the segmentation result will have both objects’ shapes fitting the occlusion boundary. The occlusion handling evaluation criterion (or method) calculates the distance between the ground truth occlusion edge and the segmentation shapes. This criterion shows how well the deep learning segmentation shape fits the ground truth occlusion edge.

This paper will be organized as follows. Following the above introduction (section 1), section 2 investigates some related works. It introduces evaluation methods for object detection and segmentation models. Section 3 shows the methodology of the evaluation criterion (or method) for measuring occlusion handling. The application of the methodology and resulting discussion are presented in section 4. Finally, section 5 presents a conclusion to this research.

2. Related works
In this paper, the authors’ have found that there is relatively little literature focusing on measuring the occlusion handling on object segmentation. In the past 10 years, different researches have employed different methods and evaluation criteria for handling occlusion in object detection. Op het Veld et al [6] proposed a method for occlusion handling. The method applied multiple classifiers, each covering a different level of occlusion and focusing on the non-occluded object parts. They evaluated their method by effecting occlusions at different object positions on the detection performance. This evaluation criterion was suitable for their method. Gao et al [7] augmented the object detection bounding box with a set of binary variables each of which corresponded to a cell indicating whether the pixels in the cell belonged to the object. This segmentation-aware representation model explicitly led to object detection more robust to occlusion. To evaluate their segmentation-aware model, they collected images including around 50% of the instances occluded by various situations. They used 5-fold cross-validation [8] to report on the model’s performance. In summary, there are various
evaluation criteria which can be used to measure the performance of different occlusion handling methods. Their evaluation criteria are unique for their special object detection methods. However, they cannot be used to evaluate the performance of occlusion handling in segmentation methods.

Another evaluation approach is to use the boundary accuracy which is suitable for object segmentation researches. Researches of segmentation methods [9, 10] have used occlusion boundary detection to enhance the object segmentation results. The recall and precision curves were used to test the detection of the occlusion boundary. Recall and precision were calculated from detected occluded pixels and ground truth occlusion boundary pixels. However, this approach particularly tests pixels wise accuracy for the occlusion boundary detection. In deep learning based segmentation, the segmented shape may not be very close to the occlusion boundary. Therefore, the boundary accuracy approach is not suitable for the evaluation of deep learning based segmentation on occlusion handling.

Another evaluation criterion can be found from [11-13]. In those researches, they computed mean IoU restricted to a narrow band of pixels around the ground-truth boundaries. This partitioning into the boundary is referred to as a tri-map in the matting literature and has been utilized in analysing semantic segmentation performance. This method can be used to test a boundary accuracy for segmentation methods. It compares the ground truth occlusion boundary and the corresponding boundary from the segmentation result. However, not all public datasets provide tri-map. The authors’ have found that only PASCAL VOC 2012 [14] and MS COCO [15] datasets provide the tri-map for testing of image segmentation. In addition, the tri-map has a parameter of Trimap Width which is the number pixels band surrounding the boundaries of the objects. The value of the Trimap Width affects the test results. Therefore, most of the researches used different Trimap Width to test their method. Thus, the test result shows a curve of accuracy vs Trimap Width [12]. Furthermore, the tri-map uses the entity object shapes to test the object segmentation methods. The dataset of VOC and COCO does not provide the tri-maps exclusively for the occlusion part. Therefore, their evaluation criterion is not an efficient approach to measure the performance of the deep learning based object segmentation on occlusion handling.

3. Methodology

In this research, the occlusion handling segmentation evaluation criterion (or method) uses the occlusion edge from the ground truth mask. A deep learning based object segmentation will segment objects. If the segmentation is perfect, the object’s segmentation (result) shape can perfectly fit the ground truth occlusion edge. The evaluation criterion is calculated from the shortest distance between the pixels on the ground truth occlusion edge and the segmentation shape. It measures how well the segmentation shape fits the ground truth occlusion edge.

If two objects are occluded with each other on an image, the ground truth mask shows an occlusion edge curve between these objects. Figure 1 (b) shows the ground truth mask image. The corresponding occlusion edge is between the two coloured masks. The ground truth occlusion edge curve is shown in Figure 1 (c), which can be obtained from the two objects’ ground truth masks by using erosion and dilation methods. The image is from the dataset of OCHuman [16] which will be explained in section 4. Suppose \( P_n(x_n, y_n) \) is a pixel location on the occlusion edge \( S_R \), if the curve has \( N \) pixels, the occlusion edge curve is defined as:

\[
S_R = \{P_n(x_n, y_n)\} = \{P_1(x_1, y_1), P_2(x_2, y_2), \ldots, P_N(x_N, y_N)\}
\]  

(1)

A deep learning based segmentation method segments the two objects on the image. The segmentation boundaries (or shapes) of the two objects may not perfectly fit the ground truth occlusion edge, but a little further on. Figure 2 shows the segmentation results of two human being objects. These results are from the instance segmentation model of Pose2Seg [16]. In Figure 2, (a) indicates a good segmentation result with masks of the two objects. The segmentation masks are almost similar to the ground truth masks. The two masks just have only a little overlapping on the left bottom corner, which causes the colour to become cyan. In Figure 2, (b) shows that the red curve is the ground truth occlusion edge. The green and blue curves are the two objects segmentation shapes. (The
three colours curves in the figure are bold. The original curves are one-pixel thickness. All the figures like this include curves with the thickness of one-pixel) In Figure 2, (c) and (d) display the ground truth occlusion edge (red) with the green and blue segmentation shapes respectively. According to Figure 2, deep learning based segmentation results do not fit the ground truth occlusion edge perfectly.

Take the blue object (object 1) segmentation result as an example. The segmentation shape \( S_p \) pixels are \( Q_m(i_m,j_m) \). If the edge has \( M \) pixels, the segmentation shape is defined as:

\[
S_p = \{ Q_m(i_m,j_m) \} = \{ Q_1(i_1,j_1), Q_2(i_2,j_2), \ldots, Q_M(i_M,j_M) \}
\]  

(2)

Figure 2. The segmentation results. (a) The segmentation result example of objects’ masks. (b) The mask shapes. The red curve is the ground truth occlusion edge. The green and blue curves are the object segmentation shapes. (c) The ground truth occlusion edge and the green object segmentation shape. (d) The ground truth occlusion edge and the blue object segmentation shape.

To measure how well the segmentation shape fits the ground truth occlusion edge, the distance between the ground truth occlusion edge and the segmentation shape should be measured. Take one pixel \( P_i(x_i, y_i) \) of the ground truth occlusion edge as an example. If the segmentation shape has \( M \) pixels, a pixel \( Q_m(i_m,j_m) \) \( (m \in [1, M]) \) from the shape has the shortest distance to the pixel \( P_i \). To explain the shortest distance, the distance calculation between two pixels on the ground truth occlusion edge and the segmentation shape is calculated as:

\[
D(P_i, Q_m) = [(x_i - i_m)^2 + (y_i - j_m)^2]^{1/2}
\]  

(3)

The shortest distance for pixel \( P_i \) on the ground truth occlusion edge to the object segmentation shape is defined as:

\[
e_1 = \min\{D(P_i, Q_m)\} = \min\{D(P_i, Q_1), (P_i, Q_2), \ldots, (P_i, Q_M)\}
\]  

(4)

The shortest distance error \( e_1 \) is the error for the \( P_i \) to the segmentation shape. This can be shown in Figure 3. Figure 3 (a) indicates that \( P_i \) is the endpoint of the ground truth occlusion (red) curve. Each pixel on the segmentation shape has a distance to \( P_i \). (a) also indicates the 5 pixels \( Q_1, Q_2, Q_3, Q_4, Q_5 \); the distances are \( D(P_i, Q_1), D(P_i, Q_2), D(P_i, Q_3), D(P_i, Q_4), \) and \( D(P_i, Q_5) \). In these five distances,
the \( D(P_1, Q_3) \) is the shortest one which is the \( e_1 \) corresponding to the \( P_1 \). Figure 3 (b) shows seven pixels from the ground truth occlusion (red) curve \((P_1, P_2, P_3, P_4, P_5, P_6, P_7)\). Their corresponding errors are \( e_1, e_2, e_3, e_4, e_5, e_6 \) and \( e_7 \). In this research, all pixels from the ground truth occlusion edge are used to calculate the shortest distance error. If the ground truth occlusion edge has \( N \) pixels, there are \( N \) errors \( \{e_1, e_2, \ldots, e_N\} \), each error is \( e_n (n \in [1, N]) \).

The distance error between the ground truth occlusion edge and the segmentation shape is the mean value of the errors \( \{e_1, e_2, \ldots, e_N\} \), which is shown below:

\[
E_1 = \frac{\sum_{n=1}^{N} e_n}{N}
\]

This distance error \( (E_1) \) shows the evaluation of the performance of the first (blue) object segmentation for occlusion handling. The green object’s (object 2) distance error \( E_2 \) is calculated using the same formula. The total occlusion handling evaluation for the segmentation is shown below:

\[
E_T = \frac{E_1 + E_2}{2}
\]

\( E_T \) is the final evaluation of the deep learning based object segmentation for occlusion handling. Because it is obtained by calculating the shortest distance between the ground truth occlusion edge and the segmentation shape, this evaluation parameter is called “Distance Error” (DE).

4. Experiment result and discussion

In this research, the Distance Error (DE) evaluation method is applied to the deep learning based segmentation model of Pose2Seg [16], who used human key points detection to enhance the occlusion handling to human segmentation. Their model was trained by using human being images from COCO dataset. They also provided a database which included heavily occluded human being images (OCHuman). In that database, 3171 images each have two human being ground truth mask labels. These images are separated into two sets: validation (1681 images) and test (1490 images). Some examples of the images are shown in Figure 4. The segmentation model of Pose2Seg was tested by this OCHuman data set, but the evaluation of the segmentation result was not clear. It only evaluated the entity human being segmentation, but it didn’t focus on the occlusion part. The DE in this research is the suitable evaluation method to show the occlusion handling for the Pose2Seg method.
Figure 4. The example of OCHuman dataset. (a) Original images. (b) The ground truth masks of two human being. (c) The ground truth occlusion edges.
Figure 4 displays examples of the OCHuman dataset and the corresponding ground truth occlusion edges. Figure 4 (a) shows the original images. Figure 4 (b) indicates the ground truth masks which have been labelled by the authors of OCHuman. Figure 4 (c) displays the ground truth occlusion edges. In this research, they are obtained from the ground truth masks (Figure 4 (b)) by using erosion and dilation methods.

In this research, the authors use the Pose2Seg model to segment human being on the images of this dataset. The results are the mask images in Figure 5. To apply the DE evaluation method, these segmentation results’ masks are used to obtain the shapes of the segmentation by using edge detection methods. After that, the equations of (3) - (6) can be applied to evaluate the Pose2Seg performance. To show the evaluation results, five images (in Figure 4) and the corresponding segmentation results (in Figure 5) are used.

The DE value indicates how well the Pose2Seg segmentation (result) shape fits the ground truth occlusion edge. The smaller the value, the better the fitting. The value of zero means the segmentation result handles the occlusion perfectly. Conversely, the bigger the value, the worse the occlusion handling of the segmentation. Figure 5 shows some examples of the DE values and the corresponding results. (The three colour curves in these two figures are bold. The original curves are one-pixel thickness.) To show how the DE value measures the occlusion handling performance, the DE value is separated into 5 level ranges.

In level 1, the DE value is less than 10. The Pose2Seg segmentation results successfully handle the occlusion on the images. Figure 5 (a)-(c) is the example of the segmentation result with the DE value of 2.98. The red curves show the ground truth occlusion edges. The blue and green curves indicate the segmentation shapes of blue and green objects. Figure 5 (b) shows that the blue object segmentation shape fits the ground truth occlusion edges well, as does Figure 5 (c).

In level 2, the DE value is between 10 and 20. The Pose2Seg segmentation shape mainly fits the ground truth occlusion. However, some small parts are not segmented well. This leads to the DE value becoming greater than 10. Figure 5 (d)-(f) indicate the example of DE value 10.03. In this example, the green object shape fits the ground truth occlusion edges well (E_2 = 4.31). Conversely, the blue object segmentation loses the crotch part, which causes the DE value to become higher (E_1 = 15.76).

In level 3, the DE value is between 20 and 30. The segmentation is worse than the level above. It loses more occlusion parts. The segmentation result loses part of the upper body or lower body. Figure 5 (g)-(i) show this example of a DE value around 20. In Figure 5 (h), the blue object segmentation shape does not fit the hands’ occlusion part. Figure 5 (i) shows that the green object segmentation loses some calf and feet region, so that the shape does not fit the occlusion of the calf and feet.

In level 4, the DE value is between 30 and 40, it shows the segmentation shapes are worse than level 3. In this situation, normally, one of the object segmentation shapes is closer to the ground truth occlusion edges but may not be perfect. However, the other object segmentation shape loses lower or upper body significantly. Figure 5 (j)-(l) indicates an example of DE value 30.15. In this segmentation result, Figure 5 (k) illustrates the blue object segmentation losing one leg and the waist. Therefore, the E_1 value is 46.22. In Figure 5 (l), the green object segmentation is too “fat”. It covers the other object’s leg and waist regions. The E_2 is 14.07.

In level 5, the DE value is greater than 40. The segmentation does not handle occlusion well. In this case, the segmentation loses most of the occluded body. DE value is sometimes unpredictable. Figure 5 (m)-(o) display an example with a DE value of 40.09. Figure 5 (n) shows the blue object segmentation shape fitting the ground truth occlusion edges approximately, but the body region is not segmented well. However, Figure 5 (o) illustrates that the green object is not segmented well. Only the head and one arm are segmented; the body part is lost. This causes the high value of DE (66.17).
Figure 5. Examples of the DE values and the corresponding segmentation results. The red curve is the ground truth occlusion edges. The blue and green colour curves are the two objects’ segmentation shapes. Each raw displays the segmentation results and the corresponding DE values relating to each raw of Figure 4 images.
Figure 5 shows the concept of the DE evaluation method for measuring the occlusion handling of Pose2Seg. Table 1 indicates the image numbers in the different range of DE values, which can show how well the Pose2Seg handles the occlusion in human being segmentation.

| OCHuman | Images No. | DE values ranges |
|---------|------------|------------------|
|         | 1 -- 10    | 10 -- 20         | 20 -- 30 | 30 -- 40 | 40 + |
| Validation | 1681        | 596              | 678      | 270      | 87   | 50   |
| Test     | 1490        | 528              | 651      | 204      | 71   | 36   |
| Total    | 3171        | 1124             | 1329     | 474      | 158  | 86   |

In Table 1, 596 images in the validation set and 528 images in the test set are in the DE range between 1 to 10. This shows about 35% of the images have successful occlusion handling results. In the DE range of 10 – 20, there are 678 in the validation set and 651 in the test set, which means 41.9% images have mainly good occlusion handling results. They just lose a small occlusion part. 474 images segmentation result in the DE value range being between 20 and 30 in total. This occupies 15% of the dataset images. These results lose more region of the body on these images. In the range of DE 30 – 40, there are 158 images. These images lose upper or lower body on the segmentation result. The DE value greater than 40 includes 86 images, which have bad results on the occlusion handling. In summary, the Pose2Seg can efficiently handle the occlusion in human being segmentation. Most of the results (almost %76.9) are in the DE value between 1 and 20.

Table 2. The mean and standard deviation of DE value.

| OCHuman | Mean DE value | Stdv DE value |
|---------|---------------|---------------|
| Validation | 15.45       | 12.89         |
| Test     | 15.02        | 10.37         |

Table 2 indicates the mean and standard deviation of DE values. They are calculated from images results. The mean values are 15.45 in the validation set and 15.02 in the test set. The standard deviation (Stdv) values are 12.89 and 10.37 in the validation and test set respectively. In the validation set, most of the result in the DE value between 2.56 (15.45 – 12.89) and 28.34 (15.45 + 12.89). The test set indicates that most of DE results are between 4.56 (15.02 – 10.37) and 25.39 (15.02 + 10.37). These values indicate that the Pose2Seg can mainly handle the occlusion efficiently.

5. Conclusion
This research introduces an evaluation method of Distance Error (DE) to measure the occlusion handling for deep learning based segmentation method. The Pose2Seg and OCHuman dataset are used to show the application of the DE method. According to DE, the Pose2Seg can efficiently handle the occlusion in human being segmentation.

The limitation of the DE method is empty segmentation. According to equation (3), this method should calculate the distances between the ground truth occlusion edge pixels and the segmentation shape pixels. However, if a deep learning model does not segment an object, the distances cannot be calculated. Future work needs to solve this limitation and make the DE method perfect.

References
[1] Fuzhen H and Xuhong Y 2010 Image segmentation under occlusion using selective shape priors *J. Image Analysis and Recognition* (Berlin Heidelberg: Springer) pp 89-95
[2] Gouda V and Banerjee S 2011 Image processing occlusion detection and handling *J. Int. Academy of Physical Sciences* 15 Special Issue 2 p 4
[3] Xiong H, Wang Z, He R and Feng D D 2012 Video object segmentation with occlusion map Int. Conf. on Digital Image Computing Techniques and Applications (DICTA) 3-5 Dec. pp 1-7

[4] Zhang W, Liu C, Chang F and Song Y 2020 Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images J. Remote Sensing 12 (no. 11) p 1760

[5] Chen Y T, Liu X and Yang M H 2015 Multi-instance object segmentation with occlusion handling Proc. IEEE Conf. on Computer Vision and Pattern Recognition pp 3470-78.

[6] Op het Veld R M, Wijnhoven R and Bondarev Y 2015 Detection and handling of occlusion in an object detection system J. Video Surveillance and Transportation Imaging Applications 9407 (International Society for Optics and Photonics) p 94070N

[7] Gao T, Packer B and Koller D 2011 A segmentation-aware object detection model with occlusion handling Conf. CVPR 20-25 June 2011 pp 1361-68

[8] Yadav S and Shukla S 2016 Analysis of k-Fold Cross-validation over Hold-out validation on colossal datasets for quality classification IEEE 6th Int. Conf. on Advanced Computing (IACC) 27-28 Feb. 2016 pp 78-83

[9] Hoiem D, Stein A N, Efros A A and Hebert M 2007 Recovering occlusion boundaries from a single image IEEE 11th Int. Conf. on Computer Vision 14-21 Oct. 2007 pp 1-8

[10] He X and Yuille A 2010 Occlusion boundary detection using pseudo-depth European Conf. Computer Vision (ECCV) (Berlin Heidelberg: Springer) pp 539-52.

[11] Ghiasi G 2016 Recognizing and segmenting objects in the presence of occlusion and clutter UC Irvine Electronic Theses and Dissertations web. https://escholarship.org/uc/item/1139s1ns

[12] Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L 2014 Semantic image segmentation with deep convolutional nets and fully connected crfs Preprint arXiv/1412.7062

[13] Kohli P and Torr P H 2009 Robust higher order potentials for enforcing label consistency Int. J. Computer Vision 82 (no. 3) pp 302-24

[14] Everingham M, Eslami S M A, Van Gool L, Williams C K I, Winn J and Zisserman A 2015 The pascal visual object classes challenge: a retrospective Int. J. Computer Vision 111 (no. 1) pp 98-136

[15] Lin T Y et al 2014 Microsoft COCO: common objects in context European Conf. on Computer Vision (ECCV) (Cham: Springer) pp 740-55.

[16] Zhang S H et al 2019 Pose2seg: detection free human instance segmentation Proc. the IEEE conf. on computer vision and pattern recognition pp 889-98