Deep Learning-Based SOLO Architecture for Re-Identification of Single Persons by Locations

Rotimi-Williams Bello, Chinedu Uchechukwu Oluigbo
Department of Mathematics and Computer Science, University of Africa, 561101 Sagbama, Bayelsa State, Nigeria

ARTICLE INFO

Article history:

Received October 17, 2022
Revised November 23, 2022
Published December 27, 2022

Keywords:

Features extraction;
Person re-identification;
SOLO;
Surveillance video;
Rank-1

ABSTRACT

Analyzing and judging of captured and retrieved images of the targets from the surveillance video cameras for person re-identification have been a herculean task for computer vision that is worth further research. Hence, re-identification of single persons by locations based on single objects by locations (SOLO) model is proposed in this paper. To achieve the re-identification goal, we based the training of the re-identification model on synchronized stochastic gradient descent (SGD). SOLO is capable of exploiting the contextual cues and segmenting individual persons by their motions. The proposed approach consists of the following steps: (1) reformulating the person instance segmentation as: (a) prediction of category and (b) mask generation tasks for each person instance, (2) dividing the input person image into a uniform grids, i.e., G×G grid cells in such a way that a grid cell can predict the category of the semantic and masks of the person instances provided the center of the person falls into the grid cell and (3) conducting person segmentation. Discriminating features of individual persons are obtained by extraction using convolution neural networks. On person re-identification Market-1501 dataset, SOLO model achieved mAP of 84.1% and 93.8% rank-1 identification rate, higher than what is achieved by other comparative algorithms such as PL-Net, SegHAN, Siamese, GoogLeNet, and M^{3}L (IBN-Net50). On person re-identification CUHK03 dataset, SOLO model achieved mAP of 82.1 % and 90.1% rank-1 identification rate, higher than what is achieved by other comparative algorithms such as PL-Net, SegHAN, Siamese, GoogLeNet, and M^{3}L (IBN-Net50). These results show that SOLO model achieves best results for person re-identification, indicating high effectiveness of the model. The research contributions are: (1) Application of synchronized stochastic gradient descent (SGD) to SOLO training for person re-identification and (2) Single objects by locations using semantic category branch and instance mask branch instead of detect-then-segment method, thereby converting person instance segmentation into a solvable problem of single-shot classification.

Corresponding Author:

Rotimi-Williams Bello, Department of Mathematics and Computer Science, University of Africa, 561101 Sagbama, Bayelsa State, Nigeria
Email: sirbrw@yahoo.com

1. INTRODUCTION

Various organizations, both public and private have recently shown interest in the use of surveillance systems. Surveillance data which includes images and videos is provided by surveillance systems as support to investigators in the investigation of crimes [1]. Moreover, the widespread use of the surveillance systems has lessened the intensity of people's fears and crimes for overall safety of the public [2]. However, it is a challenging task with a whole lot of time needed to process and analyze these images and videos for the tracking and monitoring of a person in non-overlapping cameras [3]; this problem is in addition to several other factors such as change in illumination and posture, overlapping, occlusion and complex background that negatively influence the performance of the person re-identification system in real life scenarios [4], thereby giving different appearance look to the same person. Query library and search library are the two common person re-identification tasks; while the persons of interest are contained in the query library, the images of persons which are acquired from videos by target detection algorithms are contained in the search library.

By using similarity function for matching person, the queried library image is compared with each searched library image; this process returns the person image with the highest similarity as the final recognition result [5]. In general, three critical stages are involved in the person re-identification system, namely (1) automatic person detection, (2) person features extraction and (3) classification stage. Several researchers on deep learning applications to person re-identification systems have recommended the extraction and learning of effective feature representations from the detected person body to mitigate the unwanted background objects for a robust person re-identification system [6]-[8]. This recommendation is to give a new positive dimension to the implications of learning the feature representations from the whole person image that contains unwanted background objects as found in most of the literatures. Many deep learning-based systems use convolutional neural networks (CNN) in person re-identification solution to achieve the extraction of discriminating features and feature representations due to their notable performance [4], [9]-[11]. However, to train person re-identification deep learning-based systems and achieve reliable results, a huge amount of data is required. All the aforementioned facts and issues in the existing deep learning-based systems for person re-identification arouse our interest to conduct this study.

The framework of SOLO enables the optimization of the neural network in a fashion that is end-to-end in such a way that notable limitations associated with other deep learning-based systems for person re-identification are completely removed using exclusive mask annotations for performing pixel-level solution without the need for detecting local box and grouping of pixel. In other words, the masks of each instance is directly segmented by using the annotations of a complete instance mask rather than the mask of instance present in each bounding box or the learning of affinity relations, thereby quantizing locations of the center and the sizes of the object for enablement of objects segmentation by locations. This is one of the reasons we have preferred SOLO to other mainstream segmentation models in addressing the person re-identification problem. In this paper, to achieve the re-identification goal, we based the training of the person re-identification SOLO model on synchronized stochastic gradient descent (SGD). SOLO is capable of exploiting the contextual cues and segmenting individual persons by their motions.

The proposed SOLO approach consists of the following steps: (1) reformulating the person instance segmentation as: (a) prediction of category and (b) mask generation tasks for each person instance, (2) dividing the input person image into a uniform grids, i.e., G×G grid cells in such a way that a grid cell can predict the category of the semantic and masks of the person instances provided the center of the person falls into the grid cell and (3) conducting person segmentation. Discriminating features of individual persons are obtained by extraction using convolution neural networks. The approach used in SOLO for segmentation enables it to achieve higher mAP and rank-1 identification rate results than the results achieved by PL-Net [12] and other comparative algorithms such as SegHAN, Siamese, GoogLeNet, and M^{3}L (IBN-Net50) when tested on Market-1501 and CUHK03 datasets which comprise large-scale and challenging re-identification datasets.

We applied the SOLO approach to mitigate the notable problems in the person re-identification solutions, which are the suitable methods needed for obtaining a reliable feature extraction and person features representation, and overcoming the impact of clutter environment on the identification. Moreover, different ideal solutions such as deep learning-based systems have been developed and employed by many researchers to replace the traditional manual-based feature extraction methods [13]-[16] and address the person re-identification problems with respect to feature extraction and representations. Some of these systems are in a form of a hybrid that involve two stages, namely features extraction stage and classification stage, and they are provided for identity recognition of persons across multiple cameras. Among notable works on deep learning-based systems application to person re-identification are Zhong et al. [17] who proposed a multi-level feature extraction and multi-loss learning approach as a hybrid framework using recurrent comparative network (RCN) and global average pooling (GAP) algorithms on CUHK01, CUHK03, Market-1501 and DukeMTMC-reID datasets for a better recognition of persons. By extracting multi-level attributes from different layers using feature aggregation network, a multi-level feature extraction process was achieved. Two actions are involved in the multi-loss learning process, namely verification and recognition. For the verification action, the objective is to ensure same identity of the two-person images, and the objective of the recognition action is for specific-identity of the person in each image.

A framework of optimization for learning distance metric via linear transformations was proposed by Nguyen et al. [18] by taking full advantage of the Jeffrey divergence between two multivariate Gaussian distributions resulting from local pairwise constraints. In their method, they trained the distance metric on difference spaces that are positive and negative, and constructed from the locality of each training instance, thereby preserving the local discriminative information. Wu et al. [19] focused on the person re-identification task of one-example in which there is only one labeled-example for each identity together with many examples that are unlabeled. They proposed an improved framework, which gradually utilize the unlabeled dataset for person re-identification. In their improved framework, the convolutional neural network (CNN) model was iteratively updated and pseudo labels were estimated for the unlabeled dataset.

Lu et al. [20] proposed co-attention Siamese networks for the segmentation of unsupervised video object. They achieved accurate object segmentation by effectively describing graph networks based on graph theory [21]-[22]. Based on the aforementioned facts on deep learning based methods for person re-identification, the emergence of deep learning has highly improved the performance of the person re-identification performance; although with some limitations such as insufficient training datasets for person re-identification that often causes over-fitting of the trained network model, thereby resulting in inability of person re-identification to generalize in real-life. Another limitation is inability of the extracted deep features by CNN to differentiate fine-grained person recognition effectively; these limitations and many more necessitate the need to device a new method of extracting person features with little or no influence of undesired information to achieve higher performance accuracy. The research contributions are: (1) Application of synchronized stochastic gradient descent (SGD) to SOLO training for person re-identification and (2) Single objects by locations using semantic category branch and instance mask branch instead of detect-then-segment method, thereby converting person instance segmentation into a solvable problem of single-shot classification.

2. METHODS

The methods for carrying out the proposed person re-identification system are described in this section. The system specifications include 64-bit Windows 10 Operating System, Intel Core i5 processor@2.4GHz CPU, 16 Gigabytes RAM, GeForce GTX 1080 Ti Graphics card, OpenCV Python library, Jupyter IDE, 2 Terabytes Hard-disk, 10.1 inch IPS HD Portable LCD Gaming Monitor PC display VGA HDMI interface for PS3/PS4/XBOx360/CCTV/Camera monitor. These specifications are shown in Table 1. The important parameters for training the model are: 128×128 pixels input image size, 0.0002 initial learning rate, which was later set as 0.01 for fast convergence of the model, 200 epochs for improved validation set performance, 0.0002 weight decay, 0.90 momentum, and 48-batch size.

Table 1. The software and hardware specifications

Software	**Type/Version**
Operating system	64-bit Windows 10
IDE	Jupyter
Python library	OpenCV
Hardware	**Type/Version**
CPU	Intel Core i5 processor@2.4GHz
RAM	16 Gigabytes
Graphics card	GeForce GTX 1080 Ti
Hard-disk	2 Terabytes
Monitor	10.1 inch IPS HD Portable LCD Gaming Monitor PC display VGA HDMI interface for PS3/PS4/XBOx360/CCTV/Camera

To fine-tune the hyper-parameters of the model, two large-scale public benchmark datasets were employed for the person re-identification, they are: Market-1501 dataset [23] and CUHK03 dataset [24]. Market-1501 dataset contains 1501 identities acquired from different viewpoints by 6 different cameras. The dataset contains 32,688 bounding boxes of pedestrian images that were extracted with the use of Deformable Part Models (DPM) pedestrian detector. For each person identity at each viewpoint, there are 3.6 images on average, all in .jpg format. The dataset was split into two sets of identities, with 750 identities used for training the model and 751 identities used for testing the model. Consequently, this work employed the images for the training of the proposed person re-identification system. CUHK03 dataset on the other hand, as a public dataset, contains 1,360 identities from 13,164 images in .jpg format. Campus cameras, 6 in number were deployed to capture the images, where 2 campus cameras were used to capture each identity. Two types of annotations were provided by this dataset, they are the bounding boxes that were manually labeled and the bounding boxes that were automatically detected. The dataset was split into two sets of identities in accordance with the training and testing splits described in [23], with 767 identities selected for training the model and the rest for testing the model. Table 2 shows the information on the employed person re-identification datasets.

Table 2. Information on the employed datasets

Datasets	Total number of identities	Total number of images
Market-1501	1501	32,688
CUHK03	1,360	13,164

The overall architecture of the proposed person re-identification system comprises two important stages, namely the semantic category and instance mask generation stages. Both stages are responsible for the person re-identification as shown in Fig. 1. SOLO is employed for automatic extraction of the pixel-wise mask for target persons from the complex background. During segmentation process of the input image, SOLO algorithm can perfectly handle the interference effects of the background noise and differentiate between the target object (person) and the background (undesired information) for an enhanced accurate segmentation of a person image for person re-identification solution.

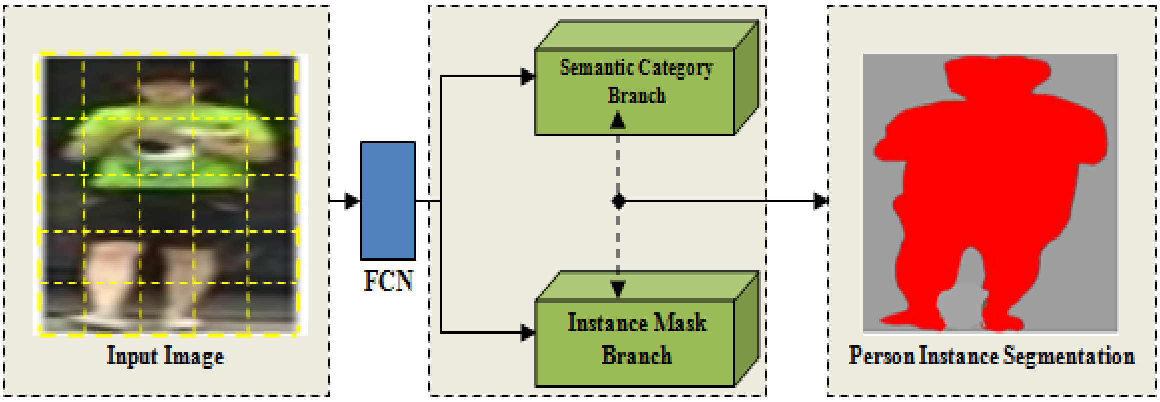

Fig. 1. The architecture of the SOLO model for addressing person re-identification problem

2.1. Semantic Category and Instance Mask Generation

For each grid, C-dimensional output (where C is the number of classes) is predicted by the proposed SOLO to show the probabilities of the semantic class which are conditioned on cells of the grid. By dividing the input person image into G×G grids, the output space becomes G×G×C. This approach is based on the premise that each G×G grid cell must contain one individual person instance, hence only containing one semantic category. During inference, the C-dimensional output shows the probability of the class for each object (person) instance. Each corresponding person instance mask is generated by each positive cell of grid in parallel with the prediction of semantic category. The fully convolutional networks (FCNs) [25] are adopted as a direct approach to predict the person instance mask, though their operations, to some extent, are spatially invariant making them more suitable for image classification for robust result.

Therefore, SOLO, being spatially invariant, is considered as the solution for the segmentation task since the segmentation masks must be conditioned on the cells of the grid and separated by different feature channels to achieve the desired result. For simple solution to the person instance segmentation task using SOLO, normalized pixel coordinates are directly fed to the networks at the initial stage of the network, and this is inspired by 'CoordConv' operator [26] for its simplicity and easy implementation. The spatial functionality is added to the FCN model by allowing the convolution access to its own input coordinates. Finally, person instance segmentation is formed based on the knowledge of naturally associating semantic category prediction and the corresponding instance mask by their reference cell of the grid. The results obtained from this segmentation are beneficial to person re-identification for obtaining parameter information about each person such as size, height, width, and length. Fig. 2 shows the system block diagrams of the two blocks of person instance segmentation process for a person re-identitication using semantic category branch and instance mask branch.

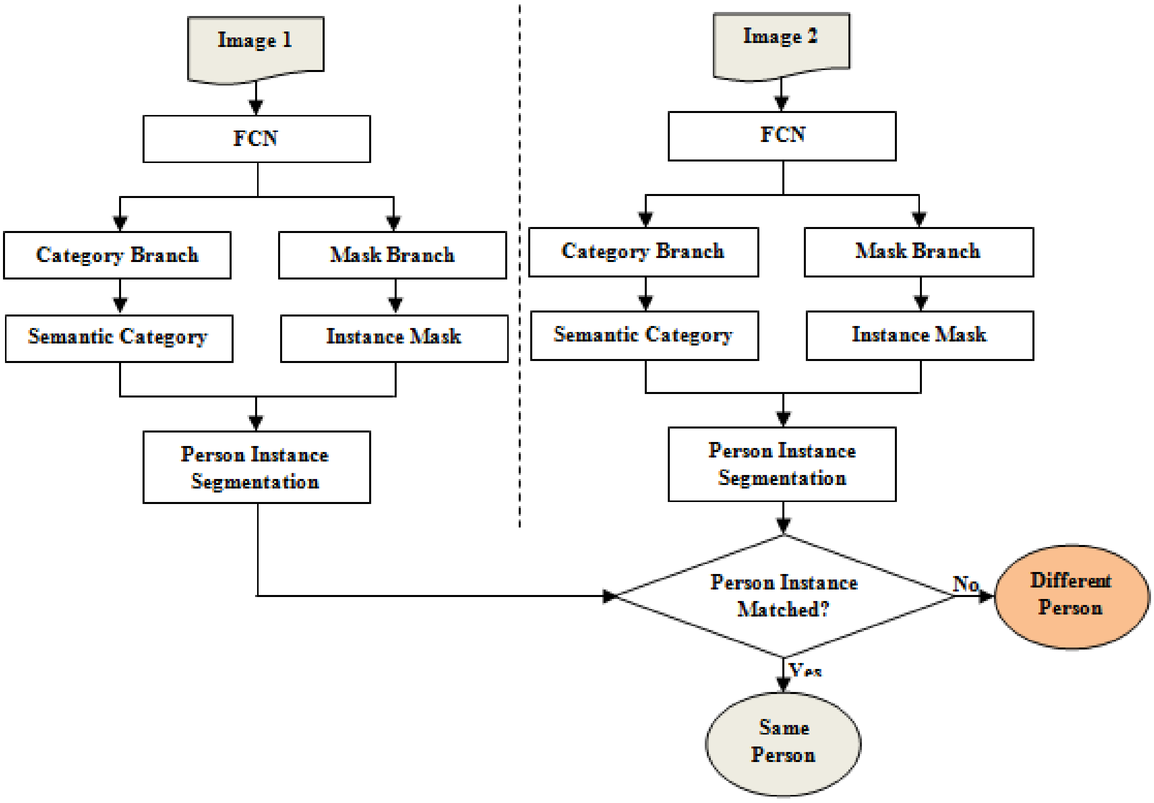

Fig. 2. System block diagrams showing two blocks of person instance segmentation process for a person re-identitication using semantic category branch and instance mask branch

2.2. SOLO Network Architecture

The construction of SOLO is such that its network is attached to the backbone of convolutional neural network (herein, for person feature extraction). Feature pyramid network (FPN) [27] is used for generating feature maps pyramid of different sizes with channels of a fixed number for each level. These maps serve as input for semantic category and instance mask as the two prediction heads, with the head weights shared across different levels, excluding only the last conv from being shared as shown in Fig. 3. SOLO architecture is instantiated with multiple architectures for its generalization and effectualness. At different pyramids, there may be variance in the number of grid, and for each person image, the differences consist of: (a) the backbone architecture employed for extracting convolutional features, (b) the network head for computing the results of the person instance segmentation, and (c) using training loss function for optimizing the model.

The head architecture handles most of the experiments (person instance segmentation) with utilization of different variants for more generalization. The generated network outputs for the person instance segmentation is shown by G = 5 grids as illustrated in Fig. 1. In Fig. 1, the right column is the person instance segmentation generated result from the networks, the left column is the input person image to the networks with G = 5 grids and the middle columns are the branches for predicting category and activating person instance masks. At each grid, only one instance is allowed for activation, however, such instance prediction may be allowed by more than two mask channels in adjacent positions. At different positions, instances are explicitly segmented so that instance (individual person) segmentation problem can be converted by SOLO into a classification task that is position-cognizant. Non-maximum suppression (NMS) is used during inference to suppress the masks that are redundant.

2.3. SOLO Learning Via Label Assignment and Loss Function

Concerning the branch that is responsible for predicting the instance category, there is a need for probability to be given by the network to the object (person) category for each of G×G grid. Explicitly, grid (i, j) which its responsibility is to indicate the generated results of mask prediction by the corresponding mask channel in the activation map is regarded as a positive sample provided the mask of ground truth has it in its center region, or else, it is regarded as a negative sample. Because there is effectualness in sampling the center region as contain in literatures on object (person) detection [29]-[37], a similar technique is utilized for classifying mask category.

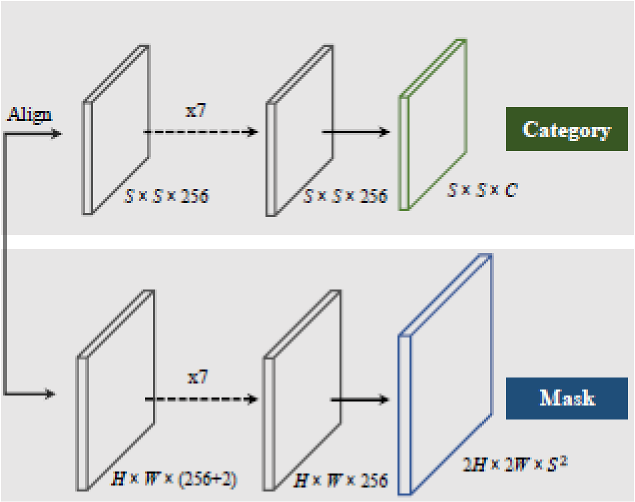

Fig. 3. Architecture of SOLO head for instance segmentation [28]. Each of the attached two sibling sub-networks at feature level of FPN is responsible for predicting instance category (top) and segmenting instance mask (bottom). This architecture is applied to person segmentation as illustrated in Fig. 1

The constant scale factors ϵ: (c_x, c_y, ϵw, ϵh) controls the center region, if the mass center (c_x, c_y), with height *h*, and width *w* of the ground truth mask are given. ϵ is set to 0.2 and there are 3 positive samples on average for each mask of the ground truth. There is a binary segmentation mask for each positive sample in addition to the label for instance category; and the binary mask of the corresponding target will be annotated for each positive sample. For each image (person image), there is S^2 output masks resulting from S^2 grids. “The loss function for the training is defined as (1)

$$L = L_{cate} + \lambda L_{mask} \tag{1}$$

where L_{cate} is the conventional focal loss [38] for semantic category classification. L_{mask} is the loss for mask prediction, and it is expressed as (2)

$$L_{mask} = \frac{1}{N_{pos}} \sum_{\mathrm{k}} 1_{(\mathbf{p}^*_{\mathbf{i,j}} > 0)}\, d_{mask}(\boldsymbol{m_k}, \boldsymbol{m^*_k}) \tag{2}$$

Here, indices *i* = [*K* / *S*], *j* = *k mod S*, if we index the grid cells (instance category labels) from left to right and top to down. N_{pos} denotes the number of positive samples, **p*** and **m*** represent category and mask target respectively. 1 is the indicator function, being 1 if $\mathbf{p}^*_{\mathbf{i,j}} > 0$ and 0 otherwise.

We have compared different implemetations of d_{mask} (.,.): binary cross entropy (BCE), focal loss and dice loss [39]. Finally, we employ dice loss for its effectivenes and stability in training. λ in (1) is set to 3. The dice loss is defined as (3)

$$L_{Dice} = 1 - D(\boldsymbol{p}, \boldsymbol{q}) \tag{3}$$

where *D* is the dice coefficient which is defined as (4)

$$D(\mathbf{p}, \mathbf{q}) = \frac{2\sum_{x,y}(\mathbf{p}_{x,y} \cdot \mathbf{q}_{x,y})}{\sum_{x,y}\mathbf{p}^2_{x,y} + \sum_{x,y}\mathbf{q}^2_{x,y}} \tag{4}$$

Here, $\mathbf{p}_{x,\,y}$ and $\mathbf{q}_{x,\,y}$ refer to the value of pixel located at (x, y) in predicted soft mask **p** and ground truth mask **q**”.

3. RESULTS AND DISCUSSION

The performance evaluation of the proposed approach which precedes the results of experiment and discussion is by using the average precision (AP) and mean average precision (mAP) being common tools for measuring and evaluating object detection and image segmentation tasks but was proposed by Zheng et al. [23] for evaluating person re-identification. The average precision and its mean are calculated as follows (5)

$$AP = \sum_{n=1}^{N} [R(n) - R(n-1) \cdot \max P(n) \tag{5}$$

where N is the calculated number of precision-recall (PR) points produced. P(n) and R(n) are the precision and recall with the lowest n-th recall, respectively.

$$mAP = \frac{1}{n} \sum_{n=1}^{N} AP_i \tag{6}$$

where AP$_i$ is the AP of class *i*, and n is the number of classes.
The equations for representing the precision rate and the recall rate are expressed as (7)

$$P = TP/TP + FP \tag{7}$$

$$R = TP/TP + FN \tag{8}$$

where, TP is true positive, FP is false positive, and FN is false negative.

Rank-k [40], another popularly and commonly employed index for evaluating the algorithm of person re-identification refers to the probability of belonging to the same queried target image, a search library image that has the first rank in similarity with the queried image. The rank-k equation is expressed as (9)

$$Rank - 1 = \frac{\sum_{i\epsilon 1,2,3,\dots,\mathrm{n}} Si}{n} \tag{9}$$

where *n* represents the number of images in their totality and S_i is an indicator that determines whether the return person image with the highest similarity belong to the same target for the final recognition when the queried library image is compared with each searched library image; (S_i returns1 if true, and S_i returns 0 if false). Rank-1 aside, other commonly employed ranks are rank-3, rank-5, rank-10 and rank-20, of which rank-1 is the simplest and the most essential index extensively employed for the performance evaluation of the model. After presenting the performance evaluation of the proposed approach, validation of the deep learning based approach of object classification in detecting the object of interest (person) is next. This approach is based on the given query image. SOLO model was examined and re-trained from scratch to suit the person re-identification datasets. By using SOLO model, there is no need for any additional model to be trained on top of the output of the segmentation procedure as found in some person re-identification work. Also, the results of the conducted SOLO experiment for the person re-identification system were qualitatively and quantitatively presented. The number of epoch (200 epochs) with which the model was trained showed best fit of the training data, and facilitated the validation set performance including a generalization ability.

The chosen image size (128×128 pixels) facilitated the speed and accuracy of the SOLO model. The rapid reduction in amount of input data was controlled by applying zero-padding on some of the input layers of the SOLO model. ResNet-101 of CNN, which is the backbone that SOLO model uses for feature extraction was used in conducting the person feature extraction on the person re-identification datasets for person re-identification solution as shown in Fig. 3. The initial configuration of the SOLO's CNN is such that each layer of the three layers used by the CNN has different number of filters which produced unsatisfactory results with the whole datasets for performance accuracy of mAP and Rank-1. To get satisfactory results, configuration adjustment was made on the second layer filter and third layer filter to increase the number of filters, whereby rate of Rank-1 and accuracy of mAP of the CNN were greatly improved on the datasets.

The results produced by our experiment on the datasets used for this study were compared to the state-of-art algorithms as shown in Table 3. On person re-identification Market-1501 dataset, SOLO model achieved mAP of 84.1% and 93.8% rank-1 identification rate, higher than what is achieved by other comparative algorithms such as PL-Net, SegHAN, Siamese, GoogLeNet, and M^{3}L (IBN-Net50). On person re-identification CUHK03 dataset, SOLO model achieved mAP of 82.1 % and 90.1% rank-1 identification rate, higher than what is achieved by other comparative algorithms such as PL-Net, SegHAN, Siamese, GoogLeNet, and M^{3}L (IBN-Net50). These results show that SOLO model achieves best results for person re-identification, indicating high effectiveness of the model. Table 3 shows the quantitative results of the feature extraction power of the SOLO model compared to other comparative algorithms such as PL-Net, SegHAN [41], Siamese [42], GoogLeNet [43], and M^{3}L (IBN-Net50) [44] and Fig. 5 shows the qualitative results from the datasets as shown in Fig. 4.

Table 3. Test results of SOLO model and other comparative algorithms on the employed datasets

Model	**Market-1501 dataset**		**CUHK03 dataset**	
	Rank-1 (%)	**mAP (%)**	**Rank-1 (%)**	**mAP (%)**
SOLO	93.8	84.1	90.1	82.1
PL-Net [12]	88.2	69.3	82.8	-
SegHAN [41]	92.3	76.1	88.3	-
Siamese [42]	83.79	74.33	50.14	50.21
GoogLeNet [43]	81.0	63.4	85.4	-
M^{3}L (IBN-Net50) [44]	75.9	50.2	33.1	32.1

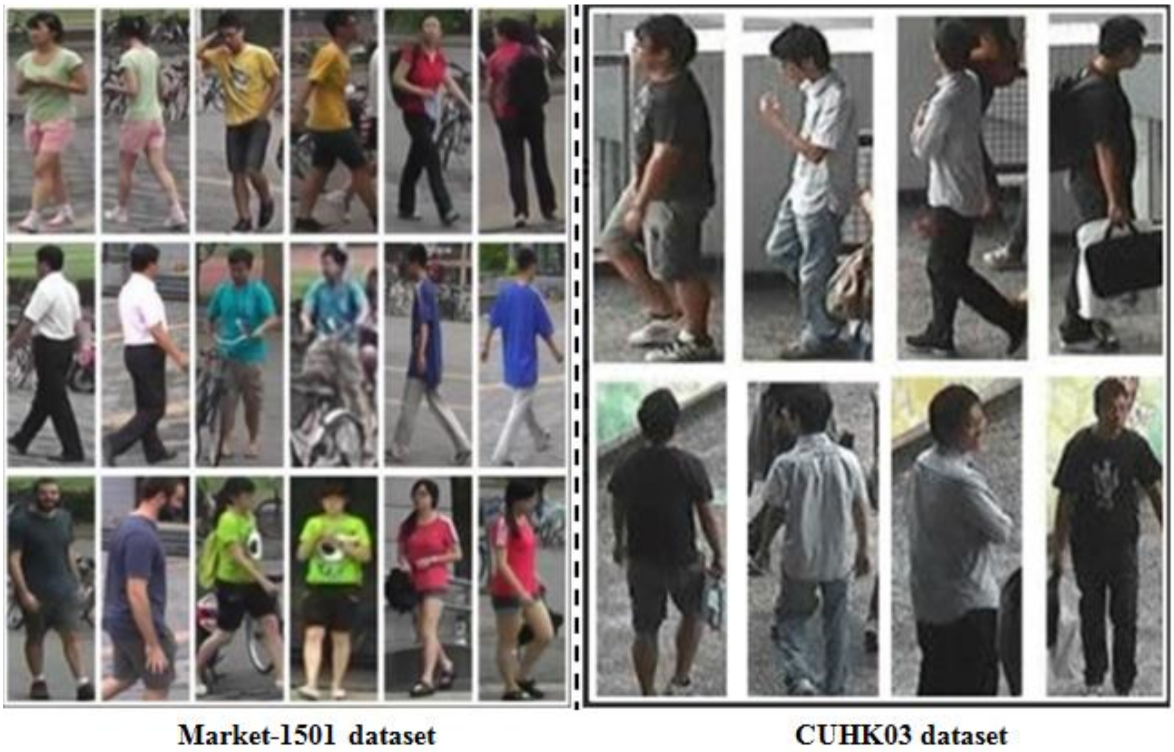

Fig. 4. Sample images from the Market-1501 and CUHK03 datasets

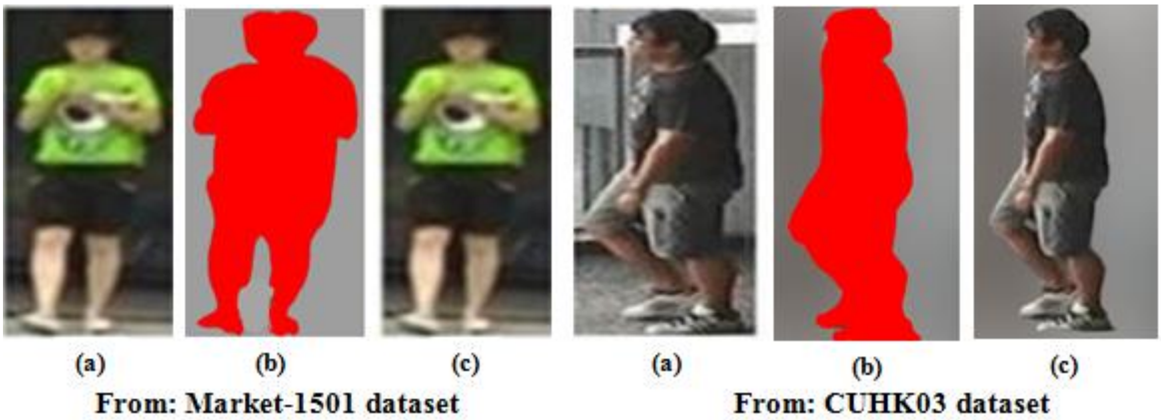

Fig. 5. SOLO segmentation procedure of person re-identification showing (a) original input images, (b) individual persons mask generation, and (c) detected individual persons. These results are based on the segmentation procedure in Fig. 1 and Fig. 2

4. CONCLUSION

A person re-identification system based on SOLO model has been proposed in this paper for matching persons at various viewpoints in non-overlapping scenes. The inability of the extracted deep features by CNN to differentiate fine-grained person recognition effectively led to using SOLO model as a method of extracting person features with little or no influence of undesired information to achieve higher performance accuracy. The construction of SOLO is such that its network is attached to the backbone of convolutional neural network. Feature pyramid network is used for generating feature maps pyramid of different sizes with channels of a fixed number for each level. Finally, person instance segmentation is formed based on the knowledge of naturally associating semantic category prediction and the corresponding instance mask by their reference cell of the grid. The results obtained from this segmentation are beneficial to person re-identification for obtaining parameter information about each person such as size, height, width and length. This parameter information can help in surveillance and security related problems. Our future research will dwell on knowing the effects of adding another trained model on top of the output of the segmentation procedure of SOLO model for person re-identification using more challenging person re-identification datasets. Moreover, as part of future work, we intend to integrate tracking algorithms to the proposed SOLO model for a robust person re-identification system.

REFERENCES

[1] D. Wu *et al.*,"Deep Learning-based Methods for Person Re-identification: A Comprehensive Review," *Neurocomputing*, vol. 337, pp. 354–371, 2019, https://doi.org/10.1016/j.neucom.2019.01.079.

[2] J. S. Kim, M. G. Kim and S. B. Pan, "A Study on Implementation of Real-Time Intelligent Video Surveillance System Based on Embedded Module," *EURASIP Journal on Image and Video Processing*, vol. 1, pp. 1-22, 2021, https://doi.org/10.1186/s13640-021-00576-0.

[3] D. Wu *et al.*, "A Novel Deep Model with Multi-Loss and Efficient Training for Person Re-identification," *Neurocomputing*, vol. 324, pp. 69–75, 2019, https://doi.org/10.1016/j.neucom.2018.03.073.

[4] M. Ye *et al.*, "Deep Learning for Person Re-identification: A Survey and Outlook," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. pp. 2872–2893, 2022, https://doi.org/10.1109/TPAMI.2021.3054775.

[5] B. Yang *et al.*, "A Feature Extraction Method for Person Re-identification Based on a Two-branch CNN," *Multimedia Tools and Applications*, vol. 81, pp. 39169–39184, 2022, https://doi.org/10.1007/s11042-022-13170-x.

[6] Z. Yao, X. Wu, Z. Xiong and Y. Ma, "A Dynamic Part-Attention Model for Person Re-identification," *Sensors*, vol. 19, no. 9, p. 2080, 2019, https://doi.org/10.3390/s19092080.

[7] R. Sun, W. Lu, Y. Zhao, J. Zhang and C. Kai, "A Novel Method for Person Re-Identification: Conditional Translated Network Based on GANs," in *IEEE Access*, vol. 8, pp. 3677-3686, 2020, https://doi.org/10.1109/ACCESS.2019.2962301.

[8] R. W. Bello, A. Z. Talib and A. S. A. Mohamed, "Real-Time Cow Detection and Identification Using Enhanced Particle Filter," *Materials Science and Engineering*, vol. 1051, no. 1, p. 01200, 2021, https://doi.org/10.1088/1757-899X/1051/1/012001.

[9] C. Ding, K. Wang, P. Wang and D. Tao, "Multi-Task Learning with Coarse Priors for Robust Part-Aware Person Re-Identification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1474-1488, 2022, https://doi.org/10.1109/TPAMI.2020.3024900.

[10] Z. Ming *et al.*, "Deep Learning-based Person Re-identification Methods: A Survey and Outlook of Recent Works," *Image and Vision Computing*, vol. 119, p. 104394, 2022, https://doi.org/10.1016/j.imavis.2022.104394.

[11] A. Genç and H. K. Ekenel, "Cross-Dataset Person Re-identification Using Deep Convolutional Neural Networks: Effects of Context and Domain Adaptation," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5843-5861, 2019, https://doi.org/10.1007/s11042-018-6409-3.

[12] H. Yao *et al.*, "Deep Representation Learning with Part Loss for Person Re-identification," in *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019, https://doi.org/10.1109/TIP.2019.2891888.

[13] W. Lin *et al.*, "Learning Correspondence Structures for Person Re-Identification," in *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2438-2453, 2017, https://doi.org/10.1109/TIP.2017.2683063.

[14] S. Li, M. Shao and Y. Fu, "Person Re-Identification by Cross-View Multi-Level Dictionary Learning," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2963-2977, 2018, https://doi.org/10.1109/TPAMI.2017.2764893.

[15] L. Wei, S. Zhang, H. Yao, W. Gao and Q. Tian, "GLAD: Global–Local-Alignment Descriptor for Scalable Person Re-Identification," in *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 986-999, 2019, https://doi.org/10.1109/TMM.2018.2870522.

[16] D. Tao, Y. Guo, B. Yu, J. Pang and Z. Yu, "Deep Multi-View Feature Learning for Person Re-Identification," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2657-2666, 2018, https://doi.org/10.1109/TCSVT.2017.2726580.

[17] W. Zhong, L. Jiang, T. Zhang, J. Ji and H. Xiong, "Combining Multilevel Feature Extraction and Multi-loss Learning for Person Re-identification," *Neurocomputing*, vol. 334, pp. 68–78, 2019, https://doi.org/10.1016/j.neucom.2019.01.005.

[18] B. Nguyen, C. Morell, and B. De Baets, "Supervised Distance Metric Learning Through Maximization of the Jeffrey Divergence," *Pattern Recognition*, vol. 64, pp. 215-225, 2017, https://doi.org/10.1016/j.patcog.2016.11.010.

[19] Y. Wu *et al.*, "Progressive Learning for Person Re-Identification with One Example," in *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872-2881, 2019, https://doi.org/10.1109/TIP.2019.2891895.

[20] X. Lu, W. Wang, J. Shen, D. Crandall and J. Luo, "Zero-Shot Video Object Segmentation With Co-Attention Siamese Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2228-2242, 2022, https://doi.org/10.1109/TPAMI.2020.3040258.

[21] Z. Zhao, S. Zhao and J. Shen, "Real-Time and Light-Weighted Unsupervised Video Object Segmentation Network," *Pattern Recognition*, vol. 120, p. 108120, 2021, https://doi.org/10.1016/j.patcog.2021.108120.

[22] T. Zhou, J. Li, S. Wang, R. Tao and J. Shen, "MATNet: Motion-Attentive Transition Network for Zero-Shot Video Object Segmentation," in *IEEE Transactions on Image Processing*, vol. 29, pp. 8326-8338, 2020, https://doi.org/10.1109/TIP.2020.3013162.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116-1124, https://doi.org/10.1109/ICCV.2015.133.

[24] W. Li, R. Zhao, T. Xiao and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152-159, https://doi.org/10.1109/CVPR.2014.27.

[25] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 2017, https://doi.org/10.1109/TPAMI.2016.2572683.

[26] R. Liu *et* al., "An Intriguing Failing of Convolutional Neural Networks and the Coordconv Solution," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1-12, 2018, https://papers.nips.cc/paper/2018/hash/60106888f8977b71e1f15db7bc9a88d1-Abstract.html.

[27] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 936-944, 2017, https://doi.org/10.1109/CVPR.2017.106.

[28] X. Wang, R. Zhang, C. Shen, T. Kong and L. Li, "SOLO: A Simple Framework for Instance Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8587-8601, 2022, https://doi.org/10.1109/TPAMI.2021.3111116.

[29] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: A Simple and Strong Anchor-Free Object Detector," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 4, pp. 1922-1933, 2022, https://doi.org/10.1109/TPAMI.2020.3032166.

[30] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li and J. Shi, "FoveaBox: Beyound Anchor-Based Object Detection," in *IEEE Transactions on Image Processing*, vol. 29, pp. 7389-7398, 2020, https://doi.org/10.1109/TIP.2020.3002345.

[31] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai and L. Zheng, "Deep Hybrid Similarity Learning for Person Re-Identification," in *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 28, no. 11, pp. 3183-3193, 2018, https://doi.org/10.1109/TCSVT.2017.2734740.

[32] X. Qian, Y. Fu, T. Xiang, Y. -G. Jiang and X. Xue, "Leader-Based Multi-Scale Attention Deep Architecture for Person Re-Identification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 371-385, 2020, https://doi.org/10.1109/TPAMI.2019.2928294.

[33] C. Zhao, K. Chen, Z. Wei, Y. Chen, D. Miao and W. Wang, "Multilevel Triplet Deep Learning Model for Person Re-identification," *Pattern Recognition Letters*, vol. 117, pp. 161-168, 2019, https://doi.org/10.1016/j.patrec.2018.04.029.

[34] J. Wang, Y. Li, S. Jiao, Z. Miao and R. Zhang, "Grafted Network for Person Re-identification," *Signal Processing: Image Communication*, vol. 80, p. 115674, 2020, https://doi.org/10.1016/j.image.2019.115674.

[35] Z. Cao and H. J. Lee, "Learning Multi-Scale Features and Batch-Normalized Global Features for Person Re-Identification," in *IEEE Access*, vol. 8, pp. 184644-184655, 2020, https://doi.org/10.1109/ACCESS.2020.3029594.

[36] N. Martinel, G. L. Foresti and C. Micheloni, "Deep Pyramidal Pooling With Attention for Person Re-Identification," in *IEEE Transactions on Image Processing*, vol. 29, pp. 7306-7316, 2020, https://doi.org/10.1109/TIP.2020.3000904.

[37] C. Yuani *et al*.,"Deep Multi-Instance Learning for End-to-End Person Re-identification," *Multimedia Tools and Applications*, vol. 77, no. 10, pp. 12437-12467, 2017, https://doi.org/10.1007/s11042-017-4896-2.

[38] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020, https://doi.org/10.1109/TPAMI.2018.2858826.

[39] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu and V. M. Patel, "KiU-Net: Overcomplete Convolutional Architectures for Biomedical Image and Volumetric Segmentation," in *IEEE Transactions on Medical Imaging,* vol. 41, no. 4, pp. 965-976, 2022, https://doi.org/10.1109/TMI.2021.3130469.

[40] R. Kat, R. Jevnisek and S. Avidan, "Matching Pixels Using Co-occurrence Statistics," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1751-1759, 2018, https://doi.org/10.1109/CVPR.2018.00188.

[41] S. Geng, M. Yu, Y. Yu and Y. Guo, "A Segmentation-Based Human Alignment Network for Person Re-identification with Frequency Weighting Re-ranking," *Academia Journal of Scientific Research*, vol. 7, no. 7, pp. 1-12, 2019, https://doi.org/10.15413/ajsr.2019.0412.

[42] C. Song, Y. Huang, W. Ouyang and L. Wang, "Mask-Guided Contrastive Attention Model for Person Re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1179-1188, 2018, https://doi.org/10.1109/CVPR.2018.00129.

[43] L. Zhao, X. Li, Y. Zhuang and J. Wang, "Deeply-Learned Part-Aligned Representations for Person Re-identification," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3239-3248, 2017, https://doi.org/10.1109/ICCV.2017.349.

[44] Y. Zhao *et al*., "Learning to Generalize Unseen Domains via Memory-based Multi-Source Meta-Learning for Person Re-Identification," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6273-6282, 2021, https://doi.org/10.1109/CVPR46437.2021.00621.