Study on Unsupervised Instance Segmentation Models for Person Re-Identification

Margarita N. Favorskaya [1], Maxim V. Savkov [1]

[1] Reshetnev Siberian State University of Science and Technology, Institute of Informatics and Telecommunications, 31, Krasnoyarsky Rabochy ave., Krasnoyarsk, 660037 Russian Federation – favorskaya@sibsau.ru, msavkov2017@gmail.com

Commission II, WG II/8

Keywords: Instance Segmentation, Unsupervised Learning, Person Re-Identification, Deep learning, Image Processing.

Abstract

Unsupervised instance segmentation for person re-identification is mainly used in challenging cases such as occluded person re-identification and 3D re-identification. Furthermore, unsupervised instance segmentation can be considered as an auxiliary cue, especially useful for long-term person re-identification using multiple cameras and single images. Several instance segmentation models, one-stage and two-stage, were examined in this study. We considered two main families of one-stage instance segmentation models: YOLO-based and SOLO-based and trained the most interesting of them. Several datasets were used for experiments, including the Market1501 dataset, the MSMT17 dataset, the DukeMTMC dataset, the DukeMTMC-reID dataset, the CUHK03 dataset, and the VIPeR dataset. The Mask R-CNN model demonstrated the best accuracy results and the YOLOACT++ model showed the best computational results in terms of instance segmentation. To compare the accuracy results without and with instance segmentation, the BUC model for person re-identification was used as a basis. The experimental results show an increase in Rank-1 accuracy values by an average of 2.7–4.9%.

1. Introduction

The problem of visual person re-identification is to search for a specific person using a query image. This formulation fundamentally distinguishes person re-identification from person tracking, which requires overlapping areas between several cameras. Biometric signals such as face and gait may be unavailable or difficult to capture in low-resolution images or video sequences. Thus, the search for a person can be carried out in different locations and at different times, which leads to four cases:

1. Short-term person re-identification using one camera and single image
2. Short-term person re-identification using one camera and short video sequence
3. Long-term person re-identification using multiple cameras and single images
4. Long-term person re-identification using multiple cameras and short video sequences

Each subsequent case removes one of the limitations and provides richer information for re-identifying the person. All these cases differ in time intervals, which involve changing or not changing clothes/wearable things (backpacks, bags, etc.) by default. Nevertheless, the problem of real-time instance segmentation must be solved in all cases. Also, person re-identification algorithms find matches between the target person image and the current images, which require unsupervised instance segmentation without the need for training datasets with manually annotated data. Moreover, instance segmentation models trained on static images can be adapted to video person segmentation through transfer learning.

Various intra-class variations such as clothes changing, occlusion, image artifacts caused by low illumination and parameters of camera, zero-shot samples, etc. make re-identification a challenging problem. According to (Jahan et al., 2024), there are three categories of unsupervised person re-identification methods, including adjusting the domain dispersion of input and output features, applying generative adversarial networks that preserve input annotations, and creating fake tags for learning by specifying similar images using clustering algorithms. One of the latest trends is the use of person's body masking (Thanh et al., 2024), for which several techniques are used such as body contour sketches, edge and shape analysis, etc. Instance segmentation based on fast deep learning models can be one of the promising ways to achieve better results.

In this study, we analyzed several unsupervised deep learning models as a potential opportunity to improve the accuracy of short-term and long-term re-identification problems using single images or short video sequences. We examine several instance segmentation models, one-stage and two-stage, with a more in-depth exploration of the one-stage models because of their speed. To illustrate the benefits of instance segmentation, we compare the accuracy results with and without instance segmentation using the latest methods, among which are the clothing agnostic shape extraction network (CASE-Net) (Li et al., 2021) and the cross-camera matching method called as "human-in-the-loop" (HITL) (Delussu et al., 2023).

The rest of the paper is organized as follows. Section 2 describes the current state-of-the-art in the field of personal re-identification. Section 3 presents a comparative analysis of unsupervised instance segmentation models. The experimental results are reported in Section 4, and, finally, conclusions are discussed in Section 5.

2. Related Work

Re-identification methods based only on visual information are addressed as appearance-based techniques, which can be divided into two groups: the learning-based methods and the direct methods. Learning-based methods are highly dependent on the type of training datasets. The underlying assumption is that the knowledge learned from the training datasets can be generalized to previously unseen samples. Direct methods are

targeted at each person independently, capturing the most distinctive aspects of colour, texture, edge, image patches, and segmented regions. However, learning-based methods produce better results than direct methods and are used more often.

Another criterion is the duration of surveillance: short-term or long-term person identification. Certainly, multiple images of the same person can provide more complete information and be employed as probe and gallery elements. It is worth noting that one of the main challenges is occlusions in real images, which leads to the problem of partial re-identification. In this study, we do not consider such cases. In any case, two fundamental options are required: robust feature extraction and similarity measure.

Since 2017, the person re-identification algorithms based on handcrafted features extraction and classification with classical machine learning methods have been rewritten for deep learning models. Among the pioneering works was the DeepDiff model, which identified the different features of various human body parts and then evaluated the similarities between them (Huang et al., 2017) and the MuDeep model as a multi-scale discriminator for cross-camera matching (Qian et al., 2017). At the same time, the unsupervised approaches have been developed, resulting in original descriptors such as multi-scale video covariance based on video tree structures (Hadjkacem et al., 2017). Additional skeletal features have become popular in the 2020s. Thus, novel metric learning method called human skeleton mutual learning person re-identification (HSMLP-Reid) was proposed in (Wang et al., 2020). This approach combined local matches based on pedestrian nodes and global skeleton matches for mutual learning.

A similarity-preserving generative adversarial network based on style transfer, consisting of a Siamese network and a CycleGAN, reduced the difference between the target domain and the source domain (Zhao et al., 2023). The ResNet-50 model was then used to extract the discriminative features common to the target and source domains. In (Yaqoob et al., 2023), the problem of resolution mismatch was solved by three steps: resizing the images, application of super-resolution network (super-resolution feedback network (Li et al., 2019)), and use of re-identification network (ResNet-50). A novel method for occluded person re-identification based on vision transformers was proposed in (Gao et al., 2024). This re-identification architecture combined point-level, part-level, and global level features to fully represent the person in holistic, self-occluded, occluded, and partial representations. In (Hambarde and Proença, 2024), a survey in soft biometric features like age, gender, clothing style, and body shape was presented as auxiliary information for person re-identification.

Re-identification raises the question of training techniques. Most existing methods typically use a ‘‘train once-and-deploy’’ scheme (domain adaptation), where identity data is collected by human annotators for offline training. However, this approach is not suitable for updating the re-identification model in real time. More and more researchers are studying the effects of ‘‘human-in-the loop’’ (unsupervised domain adaptation), which is an incremental method with humans feedback in online method based on the development of a metric matrix. Pseudo-label based methods and synthetic image generation based methods belong to unsupervised domain-adaptive methods (Delussu et al., 2023). Hard sample mining memory network combines the benefits of both deep metric learning and human-in-loop learning (Han et al., 2021). A camera-invariant and noise-tolerant framework for purely unsupervised person re-identification was suggested in (Chen et al., 2022). These authors developed a cluster memory-based meta-learning strategy using specific losses such as cluster memory-based noise-tolerant loss and a cluster memory-based additive margin loss. In (Zhang et al., 2024), a dual-stream feature fusion network (DSFF-Net) was proposed to extract discriminative RGB global features, grayscale global features, and local features from pedestrian images. However, these authors did not take advantages of semantic or instance segmentation, positioning this subtask as a future research.

This brief review shows numerous aspects for further research of the problem of personal re-identification in various problem statements.

3. Comparative Analysis of Unsupervised Instance Segmentation Models

Unsupervised instance segmentation for person re-identification can be considered as an auxiliary cue, especially useful for long-term person re-identification using multiple cameras and single images. Unsupervised instance segmentation is also applied to occluded person re-identification and 3D re-identification tasks (Ning et al., 2024).

Generally speaking, segmentation is an ill-posed problem due to differences in the perceptions of experts and hence the difficulty of obtaining ground truth data to evaluate algorithms. Traditional approaches to unsupervised instance segmentation are based on the well known K-means clustering in various colour spaces, Bayesian models, optimization methods based on energy functions, integer linear programming, superpixel-based ensemble clustering methods, and so on. However, the current focus of studies is directed on deep neural network models. And this topic is no exception.

The fundamental challenges for unsupervised instance segmentation are the following:

1. Unsupervised segmentation models are affected by the noisy pseudo-labels, resulting in a large number of false positive regions
2. Unsupervised models have weak discriminative capabilities and cannot provide comparison relations directly between different instances

There are two main categories of unsupervised instance segmentation: cluster-based and self-supervised approaches and weakly and semi-supervised instance segmentation approaches. Clustering-based instance segmentation models use the spatial relationships between points and point sets, the hierarchical aggregation using point-grouping, the graph colouring combined with fully convolutional network, among others. Self-supervised learning has a wide variety of methods, including contrastive learning, colorization, orientation discrimination, and exemplar-based approach. For example, contrastive learning can be related to partially-supervised instance segmentation method that aims to use annotated masks of base categories in order to create pseudo masks of unknown categories. The ContrastMask model, based on the Mask R-CNN model, improved feature discrimination between foreground and background regions and reduced feature dissimilarity within each region (Wang et al., 2022a). Recently, the FreeSOLO model was proposed as a self-supervised instance segmentation framework based on the SOLO model to extract coarse object masks as pseudo-labels and improve the

quality of prediction masks (Wang et al., 2022b). The Exemplar-FreeSOLO model (Ishtiak et al., 2023) is an extension of the FreeSOLO model that uses a limited number of unannotated and unsegmented exemplars.

Pixel-level annotation is computationally expensive. Thus, some instance segmentation methods use weak annotations or incomplete annotations, such as box-level annotations or image-level labels. However, such results are insufficient and worse than the results of supervised instance segmentation. Semi-supervised methods are based on the paradigm of a small number of categories with pixel-level annotations, while other categories only have block-level annotations. Thus, one promising approach is to create a good instance segmentation model from unlabeled images without any annotations.

Recently, many instance segmentation models have been developed based on deep two-stage and one-stage detectors. The simplest solution is to perform semantic segmentation followed by edge detection, pixel clustering, and creation of instance masks. This approach is the basis of the Mask R-CNN model (He et al., 2017), which is two-stage instance segmentation model: regions of interest (ROIs) are generated in the first stage and these ROIs are classified in the second stage. Mask R-CNN remains the fastest model from this group of algorithms. One-stage instance segmentation models are conceptually faster than two-stage models, but require special solutions such as position-sensitive pooling or mask voting. Zero-shot model for unsupervised object detection and instance segmentation called Cut-and-LEaRn (CutLER) was proposed in (Wang et al., 2023). The CutLER model is a two-stage model: first, it uses the MaskCut approach to create coarse masks for multiple objects in an image and, second, learns a detector on these masks using the proposed robust loss function.

Thus, since 2019, research efforts have been aimed at accelerating the assembly step of one-stage models, which was first implemented in the YOLACT (You Only Look At CoefficienTs) model (Bolya et al., 2019). The main idea of the YOLACT model based on the ResNet-101 + FPN (Feature Pyramid Network) architecture was to parallelize two tasks: creating a set of image-sized predictive "prototype masks" without depending on any instance, and predicting a vector of "mask coefficients" for each anchor followed by non-maximum suppression. The final mask for each instance was constructed by linearly combining the results of the two branches. In 2022, the same authors presented an improved model – YOLOACT++ (Bolya et al., 2022), finding a way to avoid a fully-convolutional implementation and proposing the Fast NMS (Non Maximum Suppression) algorithm. At the same time, other solutions based on the YOLO architecture were proposed. For example, the Poly-YOLO model (Hurtik et al., 2022), based on from the light SE-Darknet-53 backbone, performed instance segmentation using bounding polygons, achieving the same precision as YOLOv3 with three times smaller parameters and twice faster. The family of YOLO models involves YOLOv5-LiNet (Lawal, 2023), Poly-YOLOv8 (Zhu et al., 2023), YOLO-UNet (Iriawan et al., 2024), among others.

There are three categories of instance segmentation methods. Top-down instance segmentation methods place object instances within an a priori bounding box. Bottom-up instance segmentation methods generate instance masks by grouping the pixels into an arbitrary number of object instances. Direct instance segmentation aims to deal with instance segmentation directly without relying on box detection or embedding learning. The last concept has been implemented in the SOLO model family. The SOLO (Segment Objects by LOcations) family of models used a different paradigm (Wang et al., 2021): partitioning the input image into a spatial grid and predicting object masks using a classification score for each grid location using fully convolutional networks. The main idea of the SOLO model was to directly generate the instance masks and corresponding class probabilities, using a box-free and grouping-free paradigm. This family of models includes Vanilla SOLO, Decoupled SOLO, Dynamic SOLO (SOLOv2), Free-SOLO, Exemplar-FreeSOLO, etc. The Vanilla SOLO model had a simple architecture with two sub-networks, one for instance category prediction and one for instance mask segmentation. The Decoupled SOLO model solved the problem of sparse objects in an image when the original tensor was replaced by two tensors corresponding to two axes. The Dynamic SOLO model has been improved by using dynamic convolutions instead of traditional convolution kernels.

The FreeSOLO model was used for the task of class-agnostic instance segmentation without any annotations (Wang et al., 2022b). The FreeSOLO model was a modification of the SOLO model and contains two modules – Free Mask and Self-supervised SOLO, connected in series. The Free Mask module had a key-query attention construct that creates coarse attention masks using cosine similarity. The coarse masks were then ranked by removing redundant masks using a non-maximum suppression function. The Self-supervised SOLO module improved mask quality through a weak-supervision strategy, thereby increasing accuracy. The Exemplar-FreeSOLO model (Ishtiak et al., 2023) was an extension of the FreeSOLO model for cases with noisy pseudo-labels. The Exemplar-FreeSOLO model extended the FreeSOLO model using additional modules called the exemplar knowledge abstraction module and the exemplar embedding contrastive module. Thus, the discriminatory ability was increased through the use of exemplar guidance.

Thus, we are considering two main families of one-stage instance segmentation models – YOLO-based and SOLO-based. However, these models have not been applied for the re-identification problem. Our contribution lies in a detailed study of the possibility of their application for short-term and long-term person re-identification from single images and video sequences.

4. Experimental Results

Six datasets were used for the experiments, such as Market1501 dataset (Zheng et al., 2015), MSMT17 dataset (Wei et al., 2018), DukeMTMC dataset (Ristani et al., 2016), DukeMTMC-reID dataset (Zheng et al., 2017), CUHK03 dataset (Li et al., 2014), and VIPeR dataset (Gray et al., 2007). A brief description of each dataset is mentioned below:

1. Market1501 dataset was collected from the campus market of Tsinghua University, China, using 6 cameras, including 5 HD cameras and 1 SD camera. It includes 1,501 identities captured by six different cameras and 32,668 bounding boxes for pedestrian images. Each person has an average of 3.6 images from each viewpoint
2. MSMT17 dataset is currently the largest human Re-ID dataset, collected from 12 outdoor cameras and 3 indoor cameras. The dataset contains 126,441 pedestrian images with 4,101 identities
3. DukeMTMC (Multi-Target, Multi-Camera) dataset a dataset of surveillance videos taken on Duke University's

campus, USA in 2014. The dataset contains over 14 hours of synchronized surveillance video from 8 cameras, with over 2 million frames of 2,000 students

4. DukeMTMC-reID dataset was captured by 8 HD cameras. This dataset consists of 16,522 training images of 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images. Images were cropped by hand-drawn bounding boxes

5. CUHK03 dataset was obtained at the Chinese University of Hong Kong by two of ten cameras. The dataset includes 13,164 images with 1467 identities. The images were acquired in two ways: manual labeling of bounding boxes and a deformable part model detector for automatic detection of bounding boxes. This dataset also provides 20 random training/test splits, where 100 identifiers are selected for testing and the rest for training.

6. VIPeR (Viewpoint Invariant Pedestrian Recognition) dataset includes 632 identities captured by two outdoor cameras under different viewpoints and light conditions. Each person has one image per camera and each image has been scaled to be 128×48 pixels. It provides the pose angle of each person as 0° (front), 45°, 90° (right), 135°, and 180° (back)

Figure 1 show examples of these six datasets.

In this study, we conducted two experiments. In the first experiment, we examine the most interesting recent one-stage and two-stage models for unsupervised instance segmentation. The two-stage models include Mask R-CNN and CutLER, while the one-stage models involve YOLOACT++, Poly-YOLOv8, YOLO-UNet, SOLO, SOLO v2, and Free-SOLO.

The results of comparing 8 instance segmentation models using 5 datasets are shown in Table 1. The selected models are compared based on the following criteria:

1. IoU (Intersection over Union) is a metric used to quantify the degree of overlap between the predicted object mask and the true object mask
2. AP (Average Precision) is the average precision across all the different IoU thresholds
3. AP50 and AP75 are variations of AP where the evaluation is performed at fixed IoU thresholds of 0.50 and 0.75, respectively
4. FPS (Frames Per Second) shows how many frames per second the model can process

Table 1 does not include the DukeMTMC dataset because the DukeMTMC-reID dataset is entirely based on the DukeMTMC dataset. Moreover, the DukeMTMC-reID dataset is more specialized for solving human re-identification tasks and performs slightly better than the DukeMTMC dataset.

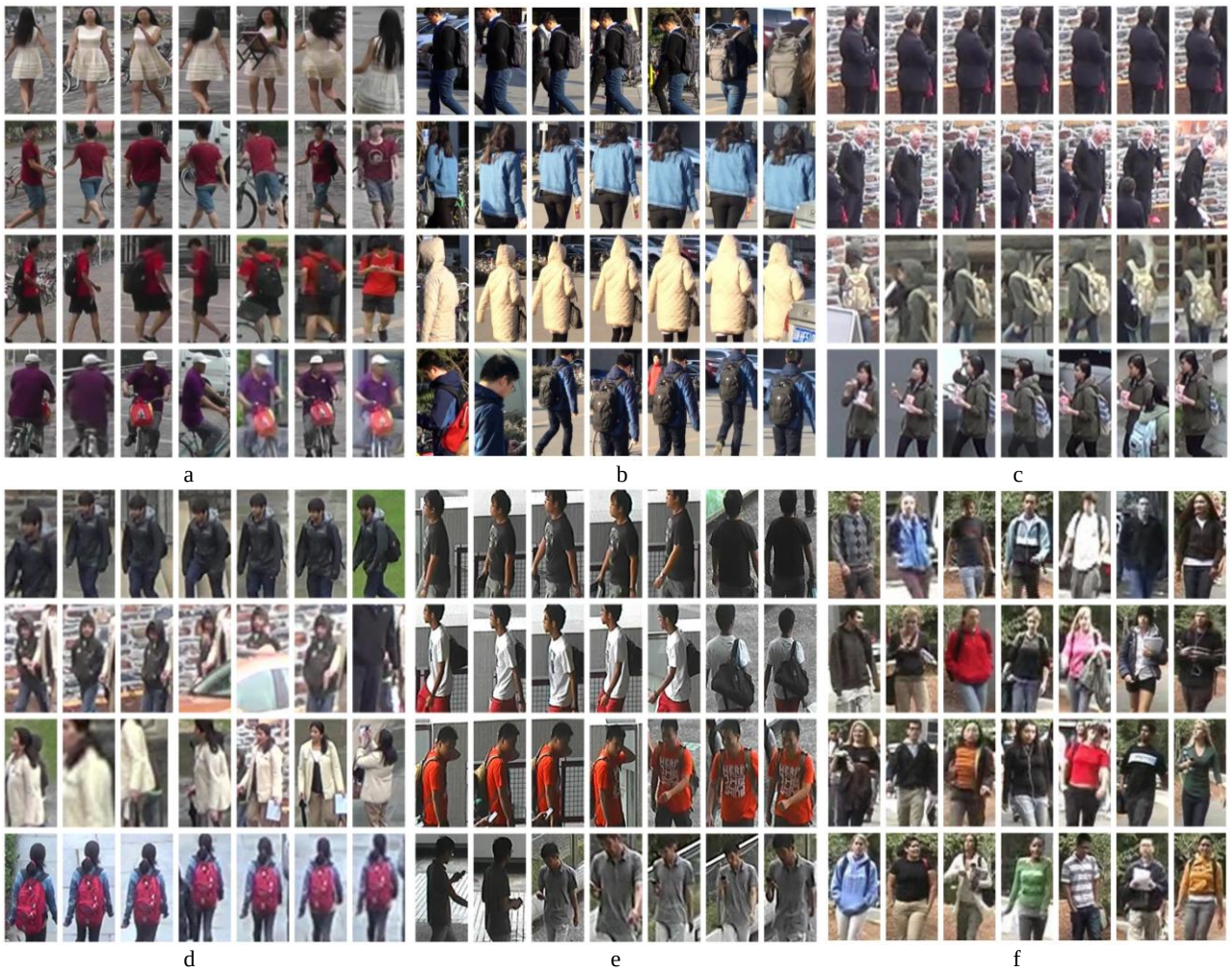


Figure 1. Sample images from various datasets: a) Market1501 dataset, b) MSMT17 dataset, c) DukeMTMC dataset, d) DukeMTMC-reID dataset, e) CUHK03 dataset, f) VIPeR dataset.

	Model	IoU (%)	AP (%)	AP_{50} (%)	AP_{75} (%)	FPS
Market-1501						
two-stage	Mask R-CNN	**77.2**	**36.1**	58.2	**37.9**	4
two-stage	CutLER	75.1	**36.1**	**58.8**	37.7	**5**
one-stage	YOLOACT++	70.2	30.2	51.5	31.1	**25**
one-stage	Poly-YOLOv8	65.3	37.2	**61.5**	39.5	20
one-stage	YOLO-UNet	71.4	33.4	56.3	35.4	10
one-stage	SOLO	66.3	34.6	57.1	35.6	9
one-stage	SOLOv2	**74.2**	38.4	59.1	38.2	12
one-stage	Free-SOLO	72.5	**40.1**	60.1	**39.2**	8
MSMT17						
two-stage	Mask R-CNN	76.3	35.1	**58.2**	**37.7**	**5**
two-stage	CutLER	**77.5**	**36.4**	**58.2**	37.1	**5**
one-stage	YOLOACT++	72.4	30.1	50.8	30.8	**26**
one-stage	Poly-YOLOv8	63.3	35.0	56.4	33.4	21
one-stage	YOLO-UNet	70.6	33.2	55.3	34.9	9
one-stage	SOLO	68.2	**37.4**	**58.5**	**38.1**	10
one-stage	SOLOv2	73.7	33.7	56.7	34.8	13
one-stage	Free-SOLO	**74.8**	33.3	53.8	33.1	9
DukeMTMC-reID						
two-stage	Mask R-CNN	**74.8**	34.2	**59.4**	36.5	3
two-stage	CutLER	74.1	**36.2**	58.6	**37.2**	**4**
one-stage	YOLOACT++	**76.1**	29.6	49.7	30.9	**22**
one-stage	Poly-YOLOv8	62.2	**35.8**	**60.9**	**38.9**	20
one-stage	YOLO-UNet	68.2	32.2	54.6	34.2	8
one-stage	SOLO	65.8	34.8	56.3	34.9	8
one-stage	SOLOv2	71.9	33.7	56.4	34.9	10
one-stage	Free-SOLO	72.7	**35.8**	55.3	35.3	11
CUHK03						
two-stage	Mask R-CNN	**72.8**	34.2	**58.4**	35.5	**4**
two-stage	CutLER	72.1	**35.2**	56.6	**36.2**	3
one-stage	YOLOACT++	**71.2**	28.6	44.7	31.9	**28**
one-stage	Poly-YOLOv8	60.3	**35.8**	**58.9**	32.9	22
one-stage	YOLO-UNet	66.4	33.2	53.6	35.2	7
one-stage	SOLO	62.5	31.8	51.3	**35.9**	9
one-stage	SOLOv2	67.8	31.7	52.4	33.9	9
one-stage	Free-SOLO	68.9	30.8	52.3	33.3	10
VIPeR						
two-stage	Mask R-CNN	**73.9**	31.9	**60.2**	**39.7**	**8**
two-stage	CutLER	73.7	**36.2**	59.6	38.2	5
one-stage	YOLOACT++	**72.5**	30.3	43.3	29.6	**30**
one-stage	Poly-YOLOv8	62.4	33.2	**57.5**	32.4	25
one-stage	YOLO-UNet	67.7	31.5	51.5	31.5	17
one-stage	SOLO	63.6	32.9	52.4	32.1	18
one-stage	SOLOv2	68.7	31.6	52.0	31.1	13
one-stage	Free-SOLO	69.8	**34.2**	56.1	**32.5**	9

Table 1. Comparison of models for 5 datasets

Table 1 shows that despite the higher accuracy of the two-stage models (Mask R-CNN and CutLER), they show significantly lower performance compared to the one-stage models (YOLOACT++, Poly-YOLOv8, YOLO-UNet, SOLO, SOLOv2, Free-SOLO). Here and further, the best results are highlihted in bold. Examples of visual instance segmentation results are depicted in Figure 2. Among the two-stage models, Mask R-CNN was chosen due to its high performance in terms of accuracy and IoU. In addition, it accurately identifies human silhouettes, even in cases of overlapping with other people or objects. Among the one-stage models, YOLOACT++ was chosen due to its good balance between accuracy and speed.

In the second experiment, we examine the impact of instance segmentation on re-identification performance by solving the problem of short-term person re-identification using one camera and single image. For this purpose, we chose one of the commonly used unsupervised models for person re-identification, called the Batch Unsupervised Clustering (BUC) model (Sculley, 2010). The BUC model for person re-identification is based on a hierarchical clustering approach. This unsupervised method iteratively merges clusters of image features, starting from individual samples and gradually building larger clusters.

In the basic configuration, BUC does not use segmentation to solve the re-identification problem. Our findings indicate that re-identification results can be improved using multitask learning, where the instance segmentation task and the re-identification task are considered as a single task with a complex loss function. Therefore, we modified the BUC model by integrating the two-stage Mask R-CNN model for instance segmentation. The Mask R-CNN model follows an algorithm that includes the following steps: sorting and filtering bindings, refining the bounding box, and generating a mask. Similarly, the one-stage YOLOACT++ model, which also implements instance segmentation, was tested. The second experiment was conducted using the Market-1501 dataset.

The accuracy results of both Mask R-CNN and YOLOACT++ are comparable, but the one-stage YOLOACT++ model demonstrates higher processing speed. Comparative precision results of the baseline BUC model, BUC with Mask R-CNN, and BUC with YOLOACT++ are presented in Table 2. Here, two metrics were used to quantify the effectiveness of the end-to-end approach. One metric is the mean Average Precision (mAP). The other one is the Rank-1 accuracy. The mAP metric reflects recall, which the Rank-1 accuracy metric reflects the retrieval precision.

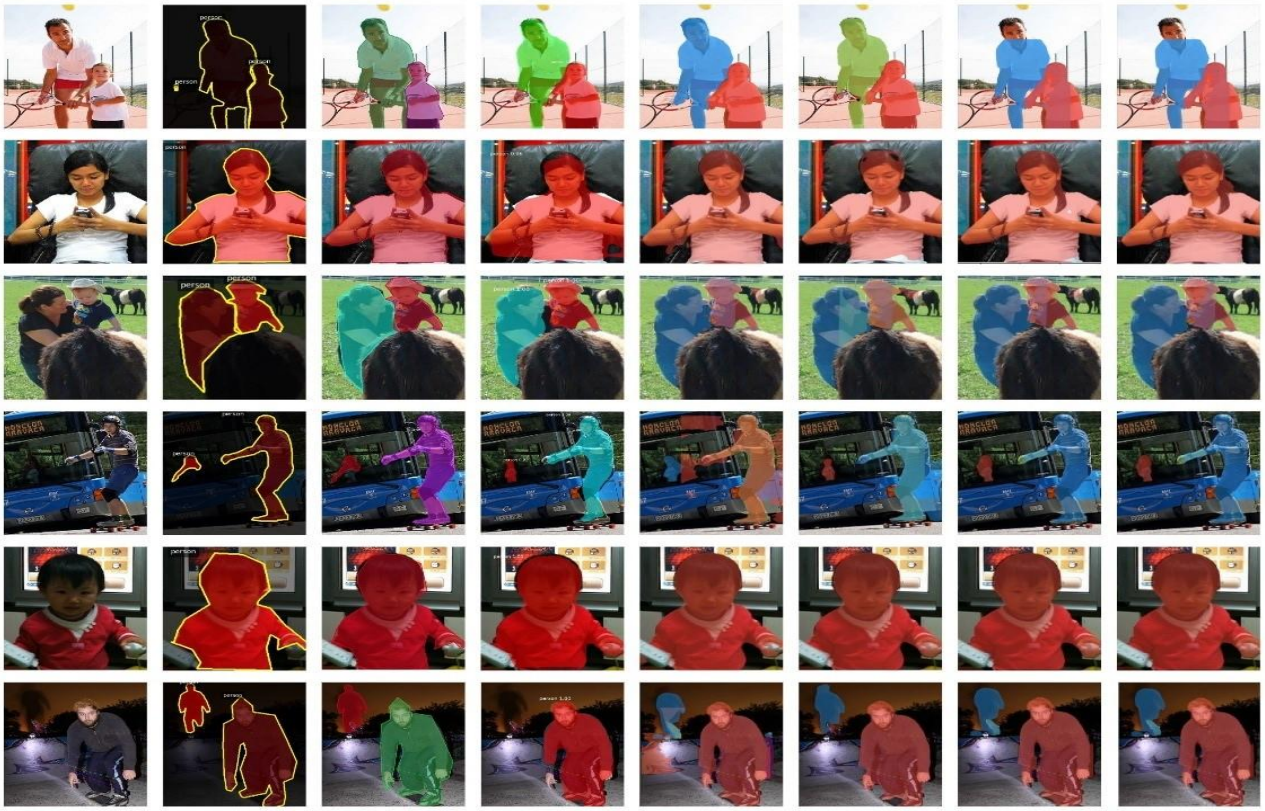
Model	mAP	Rank-1 Accuracy (%)
Baseline BUC model	0.501	68.9
BUC model with YOLOACT++	0.556	71.1
BUC model with Mask R-CNN	**0.587**	**73.8**

Table 2. Comparison of averaged precision results for baseline BUC model and modified BUC models

Table 3 provides the comparative results on the accuracy and processing time of the proposed modifications of the BUC model with the Case-NET and HITL models.

Model	Rank-1 Accuracy (%)	Processing time of 1 image (ms)
BUC with Mask R-CNN	73.8	**120**
BUC with YOLOACT++	71.1	180
Case-NET	**85.7**	210
HITL	82.3	300

Table 3. Comparison of accuracy and processing time for modified BUC models, Case-NET model and HITL model



a b c d e f g h

Figure 2. Examples of visual instance segmentation results using various models: a) orinial images, b) ground truth, c) Mask R-CNN, d) CutLER, e) YOLOACT++, f) Poly-YOLOv8, g) SOLOv2, h) Free-SOLO.

The modified BUC models, as well as other deep network models, were implemented on Python using the Pytorch repository. The GPUs used in experiment was NVIDIA Geforce RTX 3070 (8GB). The operating system is MS Windows 10.

Thus, we see that the integration of instance segmentation into the BUC model makes this modified model faster than state-of-the-art analogues with a slight decrease in accuracy. However, the proposed solution is better suited for re-identifying occluded silhouettes using instance segmentation.

5. Conclusions

Person re-identification is usually done based on a generated bounding box around the object of interest. This means that the background can influence the identification results. Multitask modification with simultaneous instance segmentation and re-identification allows to obtain more accurate results. Experiments show that the BUC model with Mask R-CNN achieves better accuracy results compared to one-stage instance segmentation models. The experiments show that the use of instance segmentation for re-identification problem is promising: although it requires additional computational costs, it ultimately makes the silhouette more accurate by reducing the influence of the background, which subsequently simplifies the person re-identification.

References

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2019. YOLACT: Real-time instance segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 9157-9166.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2022. YOLACT++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(2), 1108-1121.

Chen, Y., Fan, Z., Chen, S., 2022. Consistent camera-invariant and noise-tolerant learning for unsupervised person re-identification. *Image and Vision Computing* 123, 104462.1-104462.10.

Delussu, R., Putzu, L., Fumera, G. 2023. Human-in-the-loop cross-domain person re-identification. *Expert Systems With Applications* 226, 120216.1-120216.15.

Gao, H., Hu, C., Han, G., Mao, J., Huang, W., Guan, Q., 2024. Point-level feature learning based on vision transformer for occluded person re-identification. *Image and Vision Computing* 143, 104929.1-104929.10.

Gray, D., Brennan, S., Tao, H., 2007. Evaluating appearance models for recognition, reacquisition, and tracking. *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)* vol. 3, PETS, Rio de Janeiro, Brazil, 1-7.

Hadjkacem, B., Ayedi, W., Abid, M., Snoussi, H., 2017. Multi-shot human re-identification using a fast multi-scale video covariance descriptor. *Engineering Applications of Artificial Intelligence* 65, 60-67.

Hambarde, K., Proença, H., 2024. Image-based human re-identification: Which covariates are actually (the most) important? *Image and Vision Computing* 143, 104917.1-104917.11.

Han, P., Li, Q., Ma, C., Xu, S., Bu, S., Zhao, Y., Li, K., 2021. HMMN: Online metric learning for human re-identification via hard sample mining memory network. *Engineering Applications of Artificial Intelligence* 106, 104489.1-104489.11.

He, K., Gkioxari, G., Dollar, P., Ross Girshick, R., 2017. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2961-2969.

Huang, Y., Sheng, H., Zheng, Y., Zhang Xiong, Z., 2017. DeepDiff: Learning deep difference features on human body parts for person re-identification. *Neurocomputing* 241, 191-203

Hurtik, P., Molek, V., Hula, J., Vajgl, M., Vlasanek, P., Nejezchleba, T., 2022. Poly-YOLO: Higher speed, more precise detection and instance segmentation for YOLOv3. *Neural Computing and Applications* 34, 8275-8290.

Iriawan, N., Pravitasari, A.A., Nuraini, U.S., Nirmalasari, N.I., Azmi, T., Nasrudin, M., Fandisyah, A.F., Fithriasari, K., Purnami, S.W., Irhamah, Ferriastuti, W. 2024. YOLO-UNet architecture for detecting and segmenting the localized MRI brain tumor image. *Applied Computational Intelligence and Soft Computing* 2024, 3819801.1-3819801.14.

Ishtiak, T., En, Q., Guo, Y., 2023. Exemplar-FreeSOLO: Enhancing unsupervised instance segmentation with exemplars. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference* (CVPR) IEEE, Vancouver, BC, Canada, 15424-15433.

Jahan, M., Hassan, M., Hossin, S., Hossain, Md.I., Hasan, M., 2024. Unsupervised person Re-identification: A review of recent works. *Neurocomputing* 572, 127193.1-127193.22.

Lawal, O.M., 2023. YOLOv5-LiNet: A lightweight network for fruits instance segmentation. *PLoS ONE* 18(3), e0282297.1-e0282297.19.

Li, W., Zhao, R., Xiao, T., Wang, X., 2014. Deepreid: Deep filter pairing neural network for person re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, Columbus, OH, USA, 152-159.

Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W., 2019. Feedback network for image super-resolution. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, Long Beach, CA, USA, 3867-3876.

Li, Y.-J., Weng, X., Kitani, K.M., 2021. Learning shape representations for person re-identification under clothing change. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Waikoloa, HI, USA, 12432-2441.

Lin, Y., Dong, X,. Zheng, L., Yan, Y., Yang, Y., 2019. A bottom-up clustering approach to unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 8738-8745.

Ning, E., Wang, C., Zhang, H., Ning, X., Tiwari, P., 2024. Occluded person re-identification with deep learning: A survey and perspectives. *Expert Systems With Applications* 239, 122419.1-122419.12.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds) *Computer Vision – ECCV 2016 Workshops. ECCV 2016*, LNCS, vol. 9914, Springer, Cham, 17-35.

Qian, X., Yanwei Fu, Y., Jiang, Y.-G., Tao Xiang, T., Xiangyang Xue, X., 2017. Multi-scale deep learning architectures for person re-identification. *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, 5399-5408.

Sculley, D., 2010. Web-scale k-means clustering. *The 19th International Conference on World Wide Web*, ACM, North Carolina USA, 1177-1178.

Thanh, D.T., Lee, Y., Kang, B., 2024. Enhancing long-term person re-identification using global, local body part, and head streams. *Neurocomputing* 580, 127480.1-127480.13.

Yaqoob, I., Hassan, M.U., Niu, D., Zhao, X., Hameed, I.A., Hassan, S.-U., 2023. A novel person re-identification network to address low-resolution problem in smart city context. *ICT Express* 9, 809-814.

Wang, Z., Wei, D., Hu, X., Luo, Y., 2020. Human skeleton mutual learning for person re-identification. *Neurocomputing* 388, 309-323.

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. SOLO: A simple framework for instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(11), 8587-8601.

Wang, X., Zhao, K., Zhang, R., Ding, S., Wang, Y., Shen, W., 2022a. ContrastMask: Contrastive learning to segment every thing. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, New Orleans, LA, USA, 11604-11613.

Wang, X., Yu, Z., De Mello, S., Kautz, J., Anandkumar, A., Shen, C., Alvarez, J.M., 2022b. FreeSolo: Learning to segment objects without annotations. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA, 14176-14186.

Wang, X., Girdhar, R., Yu, S.X., Misra, I., 2023. Cut and learn for unsupervised object detection and instance segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, Vancouver, Canada, 3124-3134.

Wei, L., Zhang, S., Gao, W., Tian, Q., 2018. Person transfer GAN to bridge domain gap for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR) IEEE, Salt Lake City, UT, USA, 2018, 79-88.

Zhang, W., Li, Z., Du, H., Tong, J., Liu, Z., 2024. Dual-stream feature fusion network for person re-identification. *Engineering Applications of Artificial Intelligence* 131, 107888.1-107888.12.

Zhao, B., Wang, Y., Su, K., Ren, H., Xiyu Han, X., 2023. Semi-supervised pedestrian re-identification via a teacher–student model with similarity-preserving generative adversarial networks. *Applied Intelligence* 53, 1605-1618.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Qi., 2015. Scalable person re-identification: A benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)* IEEE, Santiago, Chile, 1116-1124.

Zheng, Z., Zheng, L., Yang, Y., 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. *2017 IEEE International Conference on Computer Vision (ICCV)* IEEE, Venice, Italy, 3774-3782.

Zhu, R., Hao, F., Dexin Ma, D., 2023. Research on polygon pest-infected leaf region detection based on YOLOv8. *Agriculture* 13(12), 2253.1-2253.17.