Vehicle Re-Identification and Tracking Based on Video Segmentation

Liangru Xiang
Department of Automation, Tsinghua University, Haidian, Beijing
xlr18@mails.tsinghua.edu.cn

Zhijia Yu
Department of Automation, Tsinghua University, Haidian, Beijing
yuzj20@mails.tsinghua.edu.cn

Jianming Hu
Department of Automation, Tsinghua University, Haidian, Beijing
hujm@mail.tsinghua.edu.cn

Yi Zhang
Department of Automation, Tsinghua University, Haidian, Beijing
zhyi@mail.tsinghua.edu.cn

ABSTRACT

Traffic object perception based on cameras is one of the foundations of Intelligent Transportation Systems. In traditional computer vision field, we usually take object detection method to detect and track the vehicle objects using bounding boxes with fixed shape, and some efficient methods based on this such as DeepSORT are used for perception. However, under the situation of dense traffic, vehicles could block each other in the viewpoint of the roadside camera, which severely reduce the accuracy of detection and tracking. Aiming to solve this, we propose our detection and tracking method based on partial feature re-identification and mask segmentation. First we apply segmentation method to separate the pixel-level image of each vehicle, then we use the especially trained CNN-based feature extractor to get the key information from the misshapen images, and finally utilize the masks and the features to track the vehicles. We test our method on CityFlow dataset and prove the validity of our method by visible result. We finally discuss the weakness of our framework and putting forward the future improvement direction of the algorithm.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**; • **Tracking**;

KEYWORDS

Vehicle re-identification, Multi-object tracking, Instance segmentation, Intelligent vehicle infrastructure cooperative system

ACM Reference Format:
Liangru Xiang, Zhijia Yu, Jianming Hu, and Yi Zhang. 2021. Vehicle Re-Identification and Tracking Based on Video Segmentation. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487185

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487185

1 INTRODUCTION

With the rapid improvement of infrastructure level, the Intelligent Vehicle Infrastructure Cooperative Systems(IVICS) has received extensive attention. In the IVICS, we expect to realize the perception, prediction and decision-making to optimize the traffic situation and avoid the accidents. Among these tasks, perception is the foundation, and how to accurately and quickly obtain the state information of traffic objects is what should be focused on. To achieve practical and instructive state information, the detection of all kinds of sensors are needed, and the method of analyzing the detection results and constructing the road network is the most important.

In the current traffic system, road detection camera is the most common traffic sensor, therefore it is necessary to develop a valid detection scheme through the camera. In the existing camera-based traffic detection framework, the object detection algorithm in computer vision is generally used to detect the positions of the vehicles in the camera field. The position is denoted by a bounding box, a rectangle with the information of x, y, width and height. In the target detection task, many fast and accurate algorithms, such as Faster-RCNN [1] and YOLO [2], have been proposed and applied in practical engineering. In addition, we want to be able to continuously track each vehicle in a video detected by the camera and maintain consistent vehicle IDs. This requires the use of multi-target tracking algorithms in computer vision. In the multi-target tracking task, there are also some excellent algorithms, such as the DeepSORT [3] algorithm, which can connect and calibrate the track of the detection box in continuous video on the basis of target detection, so as to determine the different ID of each target. The algorithm extracts the features from the photos in the detection boxes when the vehicles first appear, and use them as the features of the object for subsequent object ID comparison and re-identification.

However, when the above framework is applied to the scene of traffic, there will be a big problem: in the scene of dense vehicles, it is a very common phenomenon that vehicles occlude each other. It is conceivable that if we adopt detection strategy based on predicted boxes with those occlusions, the detection box of obscured vehicle will appear inevitably part of another vehicle. The situation is shown in Figure 1. And if use DeepSORT algorithm for feature extraction and tracking in this situation, the algorithm will wrongly add a part of other object onto the feature of the car that we focus on. And this is a disastrous result for vehicle re-identification.

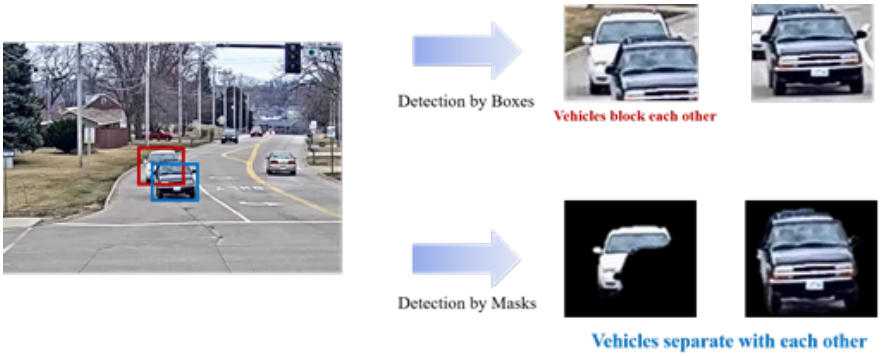

Figure 1: In Camera Viewpoints, Vehicles would Block Each Other, which Makes the Feature Vector of Each Vehicle Wrongly Include the Information of Other Object.

Figure 2: Tracking Based on Bounding Box vs Tracking Based on Segmentation Mask. Bounding Box could Get A Robust But Inprecise Notation, but the Mask Could Denote the Pixel-Level Position, and Easier to Distinguish Different Instances.

According to our experiment, the probability of ID switch was significantly higher when the vehicles occluded each other.

Considering this problem, we hope to adjust the above target detection framework, so that feature extraction under the condition of dense vehicles can exclude the influence of occlusion vehicles. Based on the DeepSORT process and the algorithm of instances segmentation, we propose an improved multi-target tracking framework, which can directly separate different targets and extract features from their exposed parts. And we extend the tracking algorithm to mask propagation based on DeepSORT algorithm. The difference between our method and traditional framework is shown in Figure 2. By making such revise, on the one hand we could simulate a more real tracking scene, on the other hand we could make some amendments to the current detection and tracking method, and promote the performance of the current method. Through experiments, our algorithm is proved to be effective to solve the occlusion problem, and could be an effective method as an addition to the bounding-box-based detection strategy.

2 RELATED WORK

2.1 Vehicle Re-identification

Vehicle re-identification is to recognize the same vehicle object, which draws more research attention in recent years. Liu et al. [4] released a high-quality VeRi-776 dataset which contains almost 50,000 pictures of 776 vehicles captured by 20 cameras. Tang et al. [5] proposed a city-scale traffic video CityFlow dataset, which provides spatial-temporal information of urban traffic. Based on those datasets, numerous methods for vehicle re-ID have been proposed. Liu et al. [6] use CNN to extract the features of vehicles, and train the model based on triplet loss and different attributes. Zhou et al. [7] utilize the Generative Adversarial Network to predict the different viewpoints of the vehicle, which could expand the dataset and promote the accuracy. Chen et al. [8] used the classification at pixel level to extract the features of different positions of the vehicle separately, so as to achieve better results in feature matching. Inspired by the part feature extraction, we design our vehicle re-ID method.

2.2 Multiple Object Tracking

Existing multi-object tracking (MOT) algorithms are usually based on target detection algorithms. The main process is as follows: target detection is carried out for each frame of the picture in a given video to obtain multiple target detection boxes; after that, visual features or motion features are used to judge the probability that the target boxes of the two frames before and after belong to the same object; finally, the detection boxes of the frames before and after are connected to form multiple target trajectories and assigned IDs. The algorithms that get the most attention in the industry are SORT [9] and DeepSORT [3]. SORT algorithm uses Kalman filter to predict and update, and Hungarian algorithm is used to solve the allocation problem, which simply realizes the

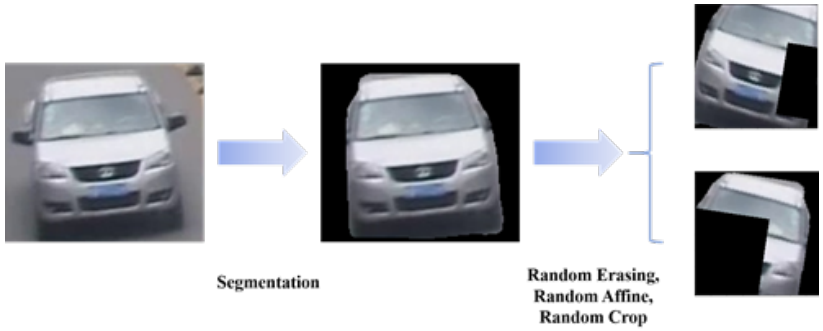

Figure 3: Data Augmentation. First, We Utilize the Mask-Rcnn to Exclude the Influence of the Background. After that We Use Random Erasing, Random Affine and Random Crop Operations to Simulate the Occluded Vehicle Pictures. By these Operations, We Could Expand the Dataset and Strengthen the Model Performance as It Detects the Blocked Cars.

framework of multi-target tracking. The DeepSORT algorithm takes it a step further by introducing appearance information into the Hungarian algorithm to solve the allocation problem, extracting features from the re-ID network and adopting the way of cascade matching, which greatly improves the tracking accuracy. Most of the other algorithms are improvements on the DeepSORT algorithm in some aspects. However, DeepSORT is an algorithm based on object detection with bounding boxes. In the case of occlusion and other problems, it is completely impossible to separate each object so that the characteristics of the vehicle can be extracted and measured independently, which makes error in the process of re-ID. In the field of pedestrian tracking and re-ID, this problem is relatively not significant, due to the human object is usually thin and always moving, which makes the blocking time short and could not influence the tracking result. However in the field of vehicle tracking, especially when detecting by road-side cameras, vehicle-block-each-other cases happen all the time, so this blocking problem is much more significant and has become the main factor which reduces the performance. Motivated by this drawback, we design our tracking method based on segmentation.

2.3 Instance Segmentation

Instance Segmentation expects to predict a mask for each object in a picture. On the one hand, it needs to finish pixel-level classification which is a part of semantic segmentation, on the other hand it needs to locate each different object in the picture even if the objects are in the same category. In computer vision, instance segmentation is one of the most difficult problem to be solved. He et al. [10] proposed an easy and efficient network Mask-RCNN for segmentation, which is the most common baseline and algorithm in this field. In our work, we take Mask-RCNN as the basic framework to segment the vehicle objects in the video.

3 OUR METHOD

Our method is divided into the following stages: partial feature-based re-recognition network, and multi-target tracking based on mask segmentation. The method is introduced as follows.

3.1 Vehicle Re-ID Based on Partial Feature

In vehicle re-recognition, the entire picture of the vehicle is generally input into the re-identification network. However, in the actual situation, the vehicles often block each other, so that the network often cannot get the complete picture of the blocked vehicle, and sometimes other objects or vehicles will be included in the picture and mix their own appearance feature into the result.

To solve this problem, we change the traditional input method and design our re-ID network based on partial feature re-identification.

Since the vehicle detected in the actual camera is likely to be occluded, the segmentation mask of the vehicle is used as the input of the re-recognition network and as the unit of target tracking. We hope that even if the picture input network is only a part of the picture of the vehicle, the output will still represent the characteristics of the vehicle. Therefore, we first perform data augmentation on the vehicle image data set.

Since we are entering an image of the vehicle behind the mask, we first take segmentation to the Veri776 dataset, extracting the vehicle portion of the original dataset and setting the rest of the unrelated background to 0. After that, we conduct random occlusion processing on the data set to randomly mask part of the information in the picture, and then use random rotation and random cropping operation, so as to simulate the exposed part of the vehicle after being occlusion. The data enhancement procession is shown as Figure 3

We set the ratio of the images conducted the enhancement procession as 0.5, which means we remain 50 percent of the data unprocessed, and 50 percent occluded, since we both need complete and occluded pictures of vehicles. We input the enhanced data into the re-identification network for training. The training loss function is divided into two parts: the first part is the cross entropy identified by the vehicle ID tag, and the accuracy is improved by minimizing the cross entropy. The formula is as follows:

$$Loss_{label} = H(P, Q) = -\sum_i P(i) \log_2 Q(i) \# \quad (1)$$

The second part is triplet loss based on metric learning. It's obvious that after data augmentation, the partial vehicle information is much less than the whole picture, the supervision of vehicle labels is not able to completely discriminate the vehicles. Therefore, we

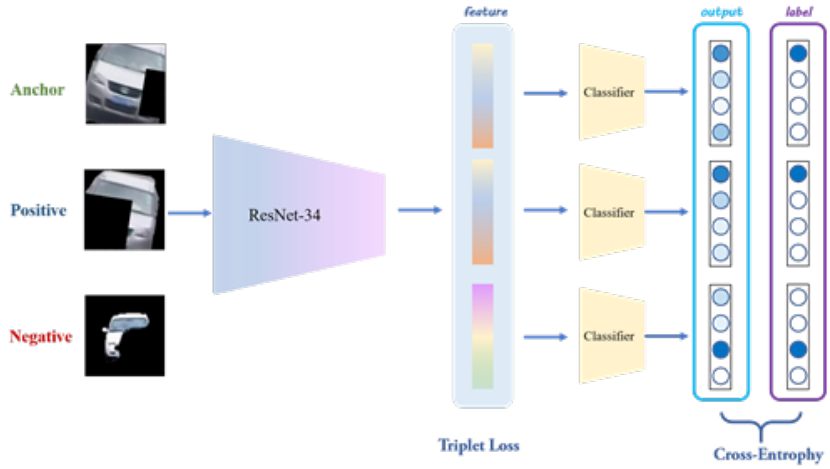

Figure 4: Total Re-ID Network. We Use ResNet-34 as Our Base Model to Extract the Car Features. Then We Use the Features to Calculate the Triplet Loss as A Part of the Loss. And the Feature Is Further Used to Identify the Vehicle ID by A Neural Network Classifier, with Cross-Entrophy to Work out the Label Loss.

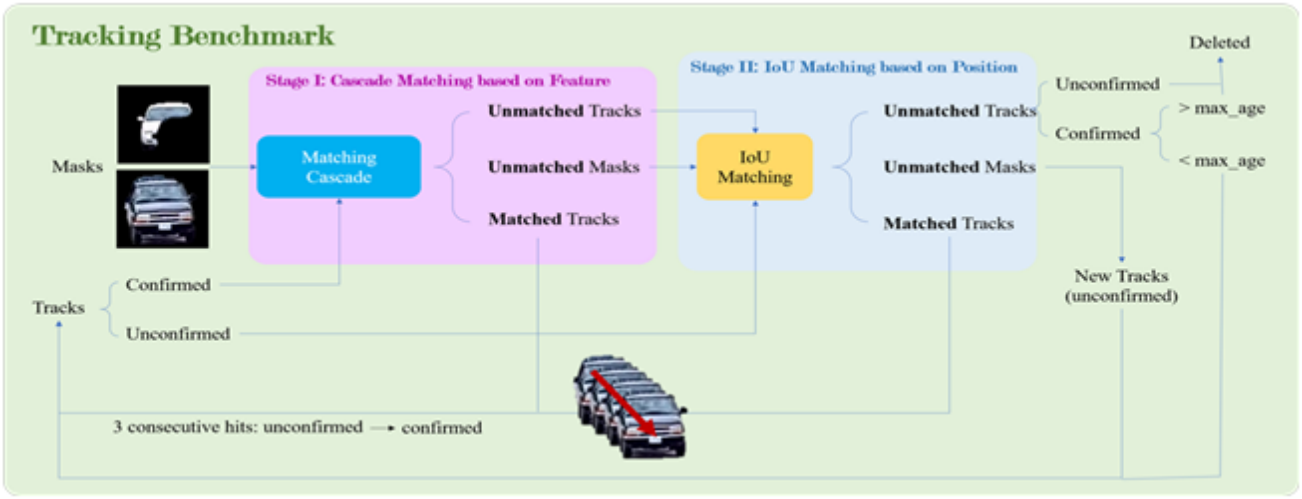

Figure 5: Multi-Vehicle Tracking Benchmark. This Procession Is Similar to DeepSORT. First We Match the Confirmed Tracks and the New-Coming Masks by Extracted Features and Prolong the Confirmed Tracks. For the Unmatched Tracks and Unmatched Masks, We Further Use IOU to Match Them. Then for those Unmatched Tracks, We Choose to Delete Them or Wait. For Those Unmatched Masks, We Choose to Create A New Track. Besides, We Confirm Those Unconfirmed Tracks if 3 Consecutive Hits Happen.

further adopt Triplet Loss[11], hoping that the feature information can still be distinguished after information reduction by means of measurement learning. The formula of triplet loss is as follows:

$$Loss_{metric} = max\left(\boldsymbol{d}\left(\boldsymbol{a},\boldsymbol{p}\right) - \boldsymbol{d}\left(\boldsymbol{a},\boldsymbol{n}\right) + \boldsymbol{margin}, 0\right)\# \quad (2)$$

The final overall loss consists of the above two parts:

$$Loss = Loss_{label} + \lambda Loss_{metric}\# \quad (3)$$

The total network achitecture is shown in Figure 4

3.2 Multi-Vehicle Tracking with Segmentation Mask

In DeepSORT, the Kalman filter of the target detection box and the Hungarian algorithm are used to generate the link of the trajectory. In our work, we use the basic structure of DeepSORT for reference, and extend the tracking target from the target detection box to the vehicle target mask.

Firstly, the mask of each vehicle object is obtained by instance segmentation of video frames with Mask-RCNN. Then the re-ID

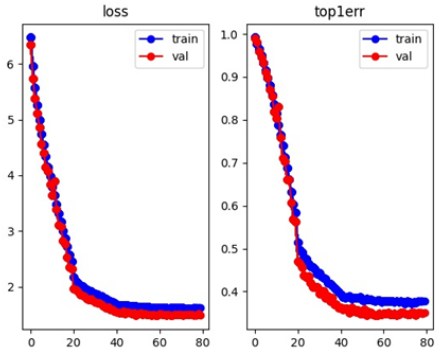

Figure 6: Training Procession. The Loss and Error on Training Set and Valiging Set Is Reducing at the Relatively Same Pace, and the Loss and Error on Training Set Is More Higher than Valiging Set Due to the Data Augmentation Applied to the Training Set.

network is used to compute these mask feature vectors corresponding to the original image. After that, we use cascade matching to match the masks and trajectories by appearance characteristics and spatial IoU. After the matching, the matched masks will be linked to the tracks, the unmatched masks would be the begin of new tracks, and the unmatched tracks would be deleted or wait according to its confirmed status. The total algorithm is similar to DeepSORT. The flow chart is shown as Figure 5

4 EXPRIMENT RESULTS

4.1 Partial Feature Re-Identification

According to the proposed segmentation and re-identification framework, we first enhance and adjust the dataset. We segment and extract the vehicle pictures from the Veri-776 Dataset by Mask-RCNN, and set the original background region in the pictures to 0. Before being input into the network, we preprocess the figures by random erasing, random affine and random crop defined in PyTorch. The randomly masked area occupies the vehicle ground truth at the ratio between 0.02 to 0.33, and the degree of random affine is limited under 30 degrees. The final random crop sets the final picture's size to 196 * 196, as the original picture size is 256 * 256. After that, we take ResNet-34 as the base model of the feature extractor, and train this model with the joint loss function derived from cross-entrophy and triplet loss. We train the network at 80 epochs, and the trend of loss and error during the training is depicted as Figure 6

After training, we test the Re-ID using the test dataset. The result is shown in Table 1

According to the test result, we could see that after data enhancement, the accuracy of network is lower than the traditional re-ID method without random erasing, which is understandable because much information of vehicles from the pictures is erased, and it's much harder to recognize the vehicle ID. However, to recognize vehicles that are blocked by other objects, our network is more efficient to extract correct and useful feature, instead of wrong information.

4.2 Multi-Vehicle Tracking

We extend the multi-target tracking algorithm to mask tracking and test our algorithm on the Cityflow dataset. We first test our method under the situation that vehicles block each other. As shown in Figure 7, although one of the two cars in the video clip blocks the other one, the mask id doesn't switch and the trajectory doesn't end. But when we use traditional DeepSORT directly, we could see that the id of the tracked vehicle changes and interrupt the trajectory that we get. Therefore we could conclude that our method is valid and more efficient than the traditional algorithm when we consider the occlusion problem.

Besides, we test the tracking result. We could see that although our method works well when vehicles block each other, the id switch of multi-object tracking sometimes happen when the vehicles do not block each other, as shown in Figure 8. This seldom happens but the ID switch is still higher than the DeepSORT. The reason is the instability of image segmentation algorithm that causes the variation of the input image and the variation of the feature. The difference of feature domain between Veri776 and Cityflow is another reason.

Due to lack of a dataset which focuses on occluded vehicles, it's hard to demonstrate the advantages of our work quantitively and directly by the current index such as IDs. And for now the result on the unblocked case performs relatively poorly than traditional method. However, our visible result could represent the validity and precision on blocked cases, which could be a guideline of future method. And a combination of our method and traditional ways is expected.

5 DISCUSSION

Although our method achieves good performance under the situation of dense traffic, it still has some defects to overcome and need further study. First of all, the tracking scheme is relatively simple, and does not fully utilize the spatio and temporal information of the video. A more efficient mask propagation method is needed. Second, the mask generated by Mask-RCNN is a little unstable for this task. According our result, many of the id-switch is caused by the shape variation of the masks. A better way to combine the bounding box and segmentation mask is expected.

Table 1: Re-ID Result on Veri776 Dataset

ACC	Rank-1	Rank-5	Rank-10
CNN without Data Enhancement	95.427%	99.443%	99.692%
Partial Feature Re-ID(Our work)	80.083%	96.055%	98.072%

Figure 7: Our Method Overcome the Block Problem, and Gain Better Detection Result than Traditional DeepSORT.

Figure 8: ID Switch Happens as Tracking by Masks. Obviously, this Is because the Lack of Stability of the Segmentation Algorithm. To Further Fix this, A Combination of Bounding Box and Segmented Mask should Be Proposed.

6 CONCLUSION

In this paper, we propose a multi-vehicle tracking framework based on video segmentation. First, we consider the blocked vehicles and train the re-ID network using data generated by a brand-new enhancement method. Second, we revise the framework of DeepSORT and extend the task to mask tracking. Finally, we summarize the feasibility and superiority of our algorithm and look forward to the future research of relative work.

ACKNOWLEDGMENTS

This work is supported by National Key R&D Program in China (2019YFF0303102) and Tsinghua University Initiative Scientific Research Program (2018Z05JDX005, 20183080016).

REFERENCES

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 91-99.
[2] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
[3] Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.
[4] Xinchen Liu, Wu Liu, Tao Mei, Huadong Ma (2016). A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. ECCV (2): 869-884
[5] Tang, Zheng and Naphade, Milind and Liu, Ming-Yu and Yang, Xiaodong and Birchfield, Stan and Wang, Shuo and Kumar, Ratnesh and Anastasiu, David and Hwang, Jenq-Neng (2019) Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8797-8806).
[6] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, Tiejun Huang (2016). Deep relative distance learning: Tell the difference between similar vehicles. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2167-2175).
[7] Zhou, Yi, and Ling Shao (2017, September). Cross-View GAN Based Vehicle Generation for Re-identification. In BMVC (Vol. 1, pp. 1-12).
[8] Tsai-Shien Chen,Chih-Ting Liu,Chih-Wei Wu,Shao-Yi Chien (2020, August). Orientation-aware vehicle re-identification with semantics-guided part attention network. In European Conference on Computer Vision (pp. 330-346). Springer, Cham.
[9] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft (2016). Simple online and realtime tracking," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 3464-3468, doi: 10.1109/ICIP.2016.7533003.
[10] K. He, G. Gkioxari, P. Dollár and R. Girshick (2020). "Mask R-CNN," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386-397, , doi: 10.1109/TPAMI.2018.2844175.
[11] F. Schroff, D. Kalenichenko and J. Philbin (2015). FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, , pp. 815-823, doi: 10.1109/CVPR.2015.7298682.