Hindawi
Journal of Sensors
Volume 2021, Article ID 9566628, 12 pages
https://doi.org/10.1155/2021/9566628

Research Article

Person Retrieval in Video Surveillance Using Deep Learning–Based Instance Segmentation

Chien-Hao Tseng,[1] **Chia-Chien Hsieh**,[1] **Dah-Jing Jwo**,[2] **Jyh-Horng Wu**,[1] **Ruey-Kai Sheu**,[3] **and Lun-Chi Chen**[3]

[1]*National Center for High-Performance Computing, National Applied Research Laboratories, Hsinchu 30076, Taiwan*
[2]*Department of Communications, Navigation and Control Engineering, National Taiwan Ocean University, Keelung 202301, Taiwan*
[3]*Department of Computer Science, Tunghai University, Taichung 40704, Taiwan*

Correspondence should be addressed to Lun-Chi Chen; lunchi@thu.edu.tw

Received 7 April 2021; Revised 27 July 2021; Accepted 29 July 2021; Published 21 August 2021

Academic Editor: Bin Gao

Video surveillance systems are deployed at many places such as airports, train stations, and malls for security and monitoring purposes. However, it is laborious to search for and retrieve persons in multicamera surveillance systems, especially with cluttered backgrounds and appearance variations among multiple cameras. To solve these problems, this paper proposes a person retrieval method that extracts the attributes of a masked image using an instance segmentation module for each object of interest. It uses attributes such as color and type of clothes to describe a person. The proposed person retrieval system involves four steps: (1) using the YOLACT++ model to perform pixelwise person segmentation, (2) conducting appearance-based attribute feature extraction using a multiple convolutional neural network classifier, (3) employing a search engine with a fundamental attribute matching approach, and (4) implementing a video summarization technique to produce a temporal abstraction of retrieved objects. Experimental results show that the proposed retrieval system can achieve effective retrieval performance and provide a quick overview of retrieved content for multicamera surveillance systems.

1. Introduction

Video surveillance systems are used in many places for security and monitoring purposes. Surveillance cameras in cities record very large amounts of video data for such purposes. Person retrieval from surveillance videos is a highly challenging task owing to camera angles, variations in illumination conditions, and multicamera surveillance systems. Moreover, manual video search is inefficient, and therefore, numerous studies have been conducted on automated video search.

The increased numbers of surveillance cameras and advances in computer vision and data storage have produced exceedingly large amounts of video sequences. Therefore, many studies have been conducted on intelligent video analysis and surveillance systems to capture information from videos and to understand the situations in scenes. Video surveillance and monitoring systems [1, 2] automatically detect and track moving objects in given video sequences. These systems help users in easily capturing information from videos. However, video sequences are long and contain large amounts of information. Therefore, these systems require considerable monitoring resources; further, they cannot easily search for past events missed in online monitoring, and these events must be identified through manual annotation.

To solve this problem, many studies have proposed event-based video retrieval systems [3–5] that can extract meaningful information or events from videos and store them in a database, enabling users to retrieve the events later. These systems can significantly reduce monitoring costs.

However, they usually focus on objects rather than on their moving patterns, and their search results are provided as still images. Therefore, it is hard to understand the motion information and high-level behaviors of objects when using these systems.

To capture the motion information of objects, this study used the concept of video abstraction (also called video summarization). Video abstraction is aimed at providing an overview of a video sequence without requiring a large amount of storage space. Video abstraction involves two main techniques: key frame extraction (also called still-image abstraction) and video skimming (also called video synopsis). Key frame extraction [6, 7] extracts the key frames that represent the original video. Many retrieval systems use the first key in a shot as the key frame. However, this method cannot analyze moving objects or dynamic events; thus, unless a video is static, this method is inefficient. In video skimming [8, 9], a long video is condensed by simultaneously presenting moving objects extracted from different periods. Therefore, this method can preserve the dynamic information of moving objects such as humans and vehicles, and the user can clearly understand the pattern of moving objects.

Recently, deep neural network-based algorithms have been widely used such as automatic ship berthing [10], person-retrieval-based pedestrian attribute recognition [11–14] owing to their efficient learning and recognition capabilities. Yang et al. presented a detection-tracking-based human action classifier that uses a convolutional neural network (CNN) and a support vector machine (SVM) based on bag-of-words to classify tracked people according to three actions of interest [15]. Geronimo and Kjellström [16] proposed an unsupervised video retrieval system that detects features of pedestrians in various scenes based on human actions and appearances. Satta [17] proposed a dissimilarity-based approach for speeding up existing reidentification methods and an approach for retrieving images of individuals based on a textual query describing clothing appearance instead of an image. Zhang et al. [18] proposed a pose-aligned neural network to recognize human attributes (e.g., age, gender, and expression) in images under unconstrained scenarios. Hu et al. [19] utilized 3D-UNet and sequence-PCA (principal component analysis) layer and designed attention block to provide enhanced semantic information of spatial and temporal multidimensional features for defect detection. The enhanced feature detecting rate provided more information in finding the defect locations. Ruan et al. [20] proposed a defect detection network (DefectNet) with a joint loss generative adversarial networks (GAN) framework for thermography defect detection. Through modifying the GAN loss and penalty loss, the detection rate of the corresponded algorithm is significantly improved. Koh et al. [21] proposed a deep temporal convolution network that addresses the need in deep learning to match the data function of a time series with an appropriate network structure. The classification accuracy of EEG and human activity signals is improved by leveraging the characteristics of the data and stabilizing the training section. However, most existing methods neglect the problem of background clutter that significantly degrades the attribute recognition and person retrieval performance. Thus, it is important to [19–21] extract more specific attribute features to capture fine-grained information from person images. In addition, Galiyawala et al. proposed a person retrieval system using a deep neural network-based approach and adaptive torso patch extraction for person detection [22]. It used mask R-CNN for semantic segmentation of a person that reduced background interference and used Tsai camera calibration [23] for height extraction. The "height" plays an important role in torso patch extraction. Torso patch extraction will remove unwanted pixels from the person image and according to adaptive "height" ratio to torso type. In this way, we must obtain the camera calibration parameters of each camera in advance. That is very inconvenient for fixed cameras.

To solve the abovementioned problems, this paper proposes a deep learning–based instance segmentation and video summarization system for video retrieval that includes an abstraction function that analyzes motion patterns and the appearance attribute features of retrieved objects. For person detection and recognition, a deep learning–based instance segmentation framework is proposed; it uses YOLACT++ [24] to perform pixelwise person segmentation to produce a masked image for each object of interest. From these contour results (masked image) of person segmentation, various human features are extracted by considering visual appearance attributes, including the color and type of clothes, to describe a person. Multi-CNN-based recognition methods are used to recognize these attributes for person retrieval. For a given query image, the retrieved objects are condensed into a shorter descriptive form using video summarization to provide a quick overview of the content retrieved from multiple camera surveillance videos. The main contributions of this paper are summarized as follows: (1) The person detection is improved by automatic instance segmentation framework to perform pixelwise person segmentation for each object of interest. It reduces noisy pixels and increases classification accuracy on a multi-CNN model. (2) A near real-time attribute recognition model is developed on the basis of appearance-based attribute features with a multi-CNN classifier. Much of the work in human appearance analysis focuses on visual-based attribute classification. However, the proposed model for attribute detection provides extracts and recognizes multiple attributes of individuals simultaneously. Its low computational cost and high accuracy are suitable for appearance attribute recognition in surveillance videos. (3) To present a briefing function via analysis of temporal information and appearance features of moving humans, this paper adopts the video summarization method to produce a temporal abstraction of retrieved objects that can present the dynamic information of retrieved objects in the video.

The remainder of this paper is organized as follows. Section 2 explains the proposed method in detail, including pixelwise person segmentation and appearance attribute feature extraction using a multi-CNN classifier. Section 3 presents the experimental results obtained using surveillance images and videos. Finally, Section 4 presents the conclusions of this study.

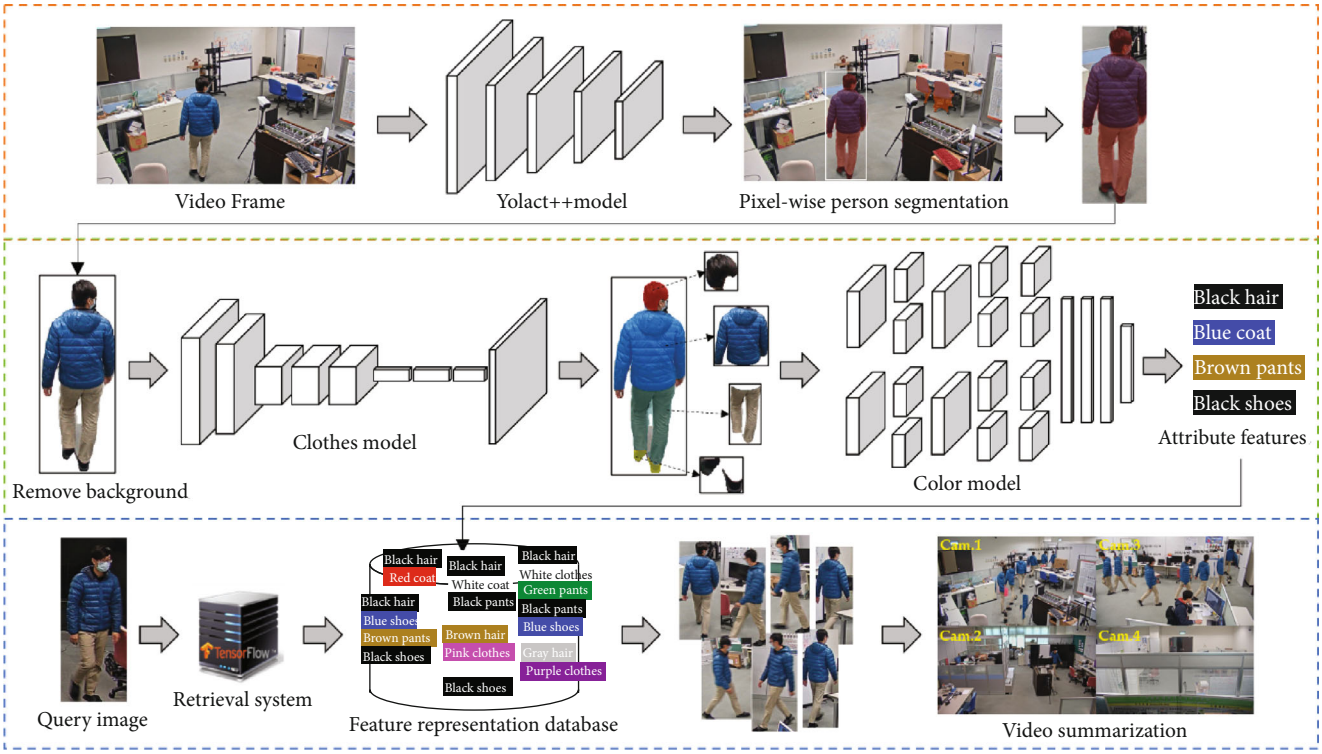


Figure 1: Overview of proposed person retrieval and summarization framework.

2. Proposed System for Person Retrieval and Summarization

Figure 1 presents the framework of the proposed deep learning–based person retrieval system. A pixelwise instance segmentation network is employed to detect the person (foreground) region and remove the background. Further, a multi-CNN classifier is used to extract the appearance attribute features of a human masked image to represent a person. Lastly, a video summarization function is developed to provide a quick overview of the content retrieved from multicamera surveillance videos.

2.1. Pixelwise Person Segmentation. Moving object detection is the first step to detect instances of semantic objects of a certain class, such as humans, animals, or cars, in a sequence of videos. Object detection can be performed through various approaches such as frame-to-frame difference, background subtraction, and motion analysis using optical flow techniques. These traditional approaches typically use feature selection and extraction algorithms to recognize instances of an object category. Traditional moving object detection methods are divided into two categories: moving object detection, which includes background subtraction, optical flow, and spatiotemporal filtering, and object classification and recognition, which primarily uses shape-based, motion-based, and texture-based visual features. However, these traditional methods are easily affected by external factors in actual operation, resulting in low accuracy of the detection results.

Recent years have seen significant improvements in moving object detection owing to the development of deep learning–based methods. Such methods not only improve the accuracy of object detection but also enable object classification and semantic segmentation in the same neural network model, called an instance segmentation model. Many studies have been conducted to improve object segmentation accuracy. Mask R-CNN [25] is a representative two-stage instance segmentation model that first generates candidate regions of interest (ROIs) and then classifies and segments these ROIs in the second stage. Follow-up studies improved its accuracy by enriching the feature pyramid network (FPN) features [26] and addressing the incompatibility between the confidence score and the localization accuracy of a mask [27]. These two-stage methods require repooling features for each ROI and processing them with subsequent computations; this makes them unable to obtain real-time speeds (30 fps) even if the image size is reduced.

One-stage instance segmentation methods generate position-sensitive maps that are assembled into final masks with position-sensitive pooling [28, 29] or combine semantic segmentation logits and direction prediction logits [30]. Although they are conceptually faster than two-stage methods, they still require repooling or other nontrivial computations. This severely limits them to subreal-time speeds. Some methods first perform semantic segmentation followed by boundary detection [31], pixel clustering [32, 33], or learning an embedding to form instance masks [34–37]. However, these methods involve multiple stages and/or expensive clustering procedures, thus limiting their viability for real-time applications.

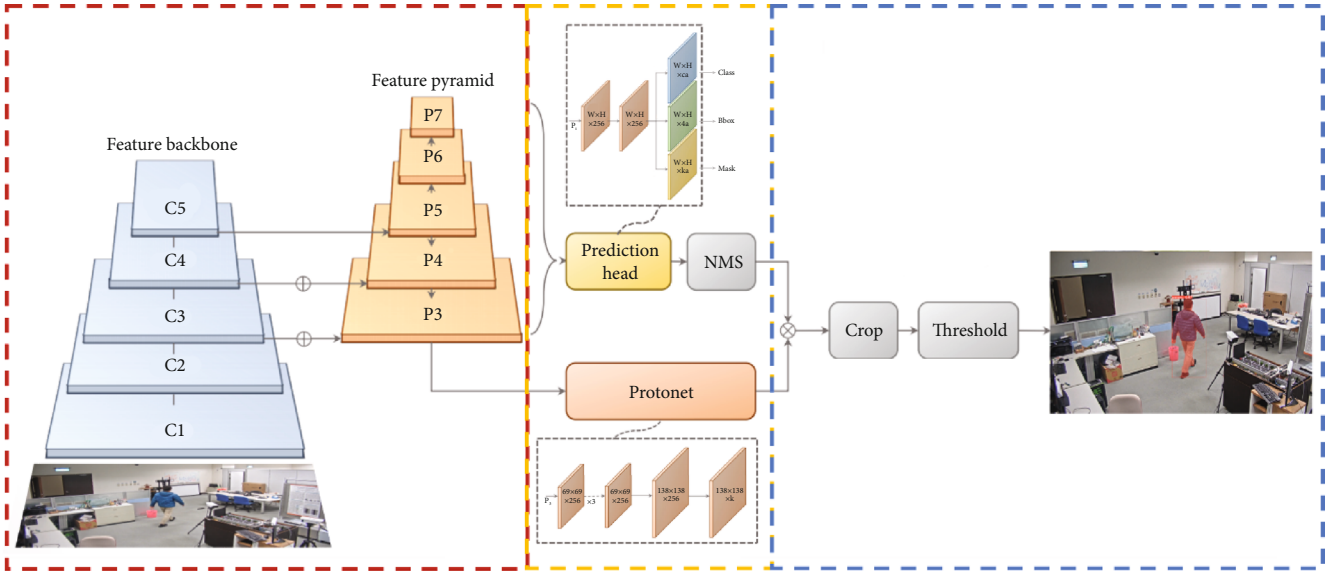


Figure 2: Architecture of YOLACT++ model. It uses ResNet-101 with FPN as the feature backbone, and it contains a classification structure (top branch) and a semantic segmentation structure (bottom branch) for the entire image.

In this study, we applied the state-of-the-art YOLACT++ [25] model to achieve real-time (>30 fps) instance segmentation. Experimental results indicate that YOLACT++ is significantly faster than many previous models. The framework of the YOLACT++ model divides instance segmentation into two parallel subtasks: (1) generating a set of prototype masks and (2) predicting per-instance mask coefficients. Figure 2 depicts the architecture of the YOLACT++ model.

They produce instance masks by linearly combining the prototypes with the mask coefficients. Because this process does not depend on repooling, this approach produces especially high-quality masks and also exhibits temporal stability. Fast nonmaximum suppression (NMS), a drop-in replacement for standard NMS that is 12 ms faster and has only a marginal performance penalty was also proposed. Finally, by incorporating deformable convolutions into the backbone network, optimizing the prediction head with better anchor scales and aspect ratios, and adding a novel fast mask rescoring branch, the YOLACT++ model can achieve 34.1 mAP on the MS Common Object in Context (COCO) dataset at 33.5 fps, which is fairly close to previous state-of-the-art approaches, while running in real time. In our experiments, we used a ResNet-101 model that is pretrained using the MS COCO dataset and FPN as our default feature backbone.

2.2. Appearance Attribute Feature Extraction Using Multi-CNN Classifier. In a video surveillance system, the ability to recognize the identity of a certain person in the video sequence is vital. To identify a certain person amidst various spatial and temporal video information, we need to record sufficient information for each human segmentation. Describing clothing is a straightforward method to identify a person because the combinations of various colors and types of clothes uniquely identify persons.

Feature extraction methods are generally divided into two categories: traditional keypoint descriptor approaches such as SURF [38] and SIFT [39] and CNN-based feature extractors like CNN [40]. Compared with handcrafted features, CNN-based feature extractors usually perform better because they can robustly extract features using deep learning to jointly extract discriminant object features for classification. Generally, most functional CNN-based feature extractors are trained on different types of objects that may have prominent distinguishing features. However, this study is aimed at distinguishing the differences within color classes of people's clothes. To deal with this problem, we needed to retrain the CNN model, and more importantly, create more discriminant features for different classes of clothing colors in multicamera surveillance networks.

The model functions are as follows. To reduce noise, each image of a video sequence is input to the YOLACT++ model for detection and instance segmentation of people. The unmasked area is zero-filled if the detected object is classified as a person and its confidence score is above a certain threshold. Then, a multi-CNN classifier is used to extract the appearance features of a human masked image to represent a person. Therefore, we use an appearance-based attribute feature method with a multi-CNN pipeline involving two steps: (1) appearance feature extraction and (2) color classification.

2.2.1. Appearance Feature Extraction. We pretrain a fully convolutional network (FCN) with the Look into Person (LIP) [41] dataset that contains a total of 10 classes of clothing data: hat, hair, glove, clothes, coat, socks, pants, scarf, skirt, and shoes, shown in Figure 3. In this study, we adopted an FCN model to extract people's appearance features through two steps. First, the model mentioned in Section 2.1 is used to perform pixelwise instance segmentation of people. Because the size of the person segmentation images may be different, we resize all images such that they can be used as an input of the FCN model. Second, we split the

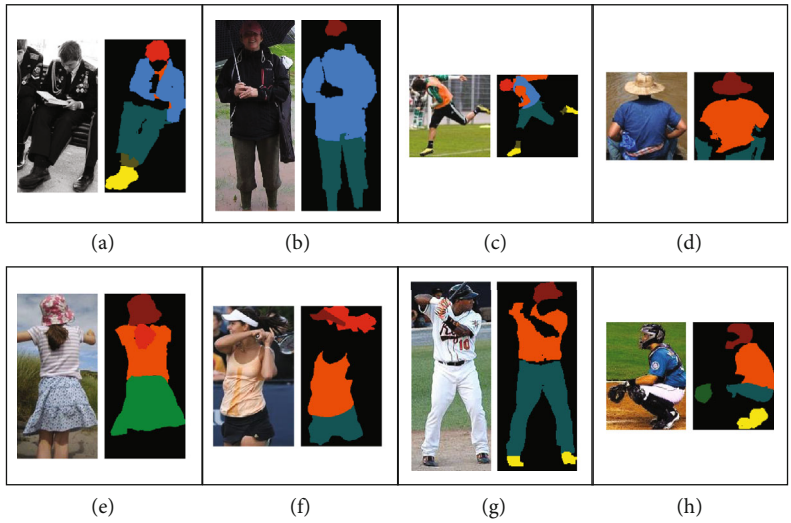


Figure 3: Sample images of the LIP dataset.

person segmentation images into multiple appearance features, such as hair, pants, jacket, and shoes, using this FCN model.

2.2.2. Color Classification. We use a CNN classifier to recognize the color of every appearance feature by retraining the CNN model using our multicamera video data. The architecture of the CNN model consists of two base networks with eight layers in each base network for a total of 16 layers. The color recognition model has been described previously [40]. In this study, we used real data; Figure 4 shows some sample images from the dataset. The dataset contains images of clothes with 11 classes of color: black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow.

At the end of this pipeline, we can automatically extract multiple appearance attribute features with sufficient information for the representation of each person. Figure 5 shows the extraction of appearance attributes and color features.

2.3. Person Retrieval and Summarization. Person retrieval is aimed at retrieving the images of a specified pedestrian in a large database in response to a queried person of interest. The proposed system provides the event retrieval results (e.g., "blue" + "coat") as dynamic videos, whereas traditional event retrieval systems usually provide still images. However, if the number of the retrieved events is large, watching the resulting videos may be time consuming. Therefore, the proposed system performs video summarization to condense a long video while maintaining the information of moving objects in the video.

We first use Resnet-101 in the YOLACT++ model to detect and segment persons in a multicamera environment. The segmented persons are retrieved through the retrained FCN model and color model. The retrieved features are compared with an attribute feature representation database to determine the identity of the person in the image. Section 2.2 already described the method for training the FCN model and the retrieved features. Although we can recognize the

Figure 4: Sample images of clothes of different color classes.

identity of the key person in video sequences, the key person may appear in the image without being noticed when shooting for long durations. In such cases, information from images of the key person must not be missed at each moment. We collect information of the key person in the image at each moment and project it in the abstraction image. The abstraction image will show the position of the key person in the image at each moment. Figure 6 shows event abstraction for producing a condensed video. Figure 7 shows the video summarization process in a multicamera surveillance system.

3. Experiment Results

In the previous section, we introduced the related methods and processes used in our proposed system. This section presents an evaluation of the performance of the proposed

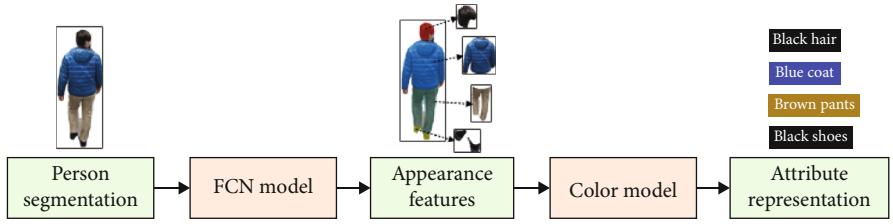


Figure 5: Extraction of appearance attributes and color features.

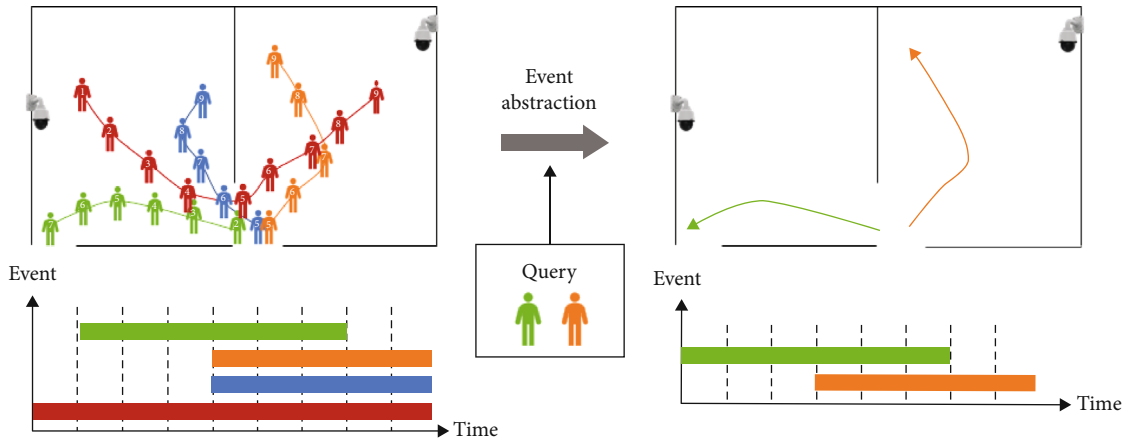


Figure 6: Illustration of event abstraction. Selected persons of interest are temporally condensed according to the query attribute features.

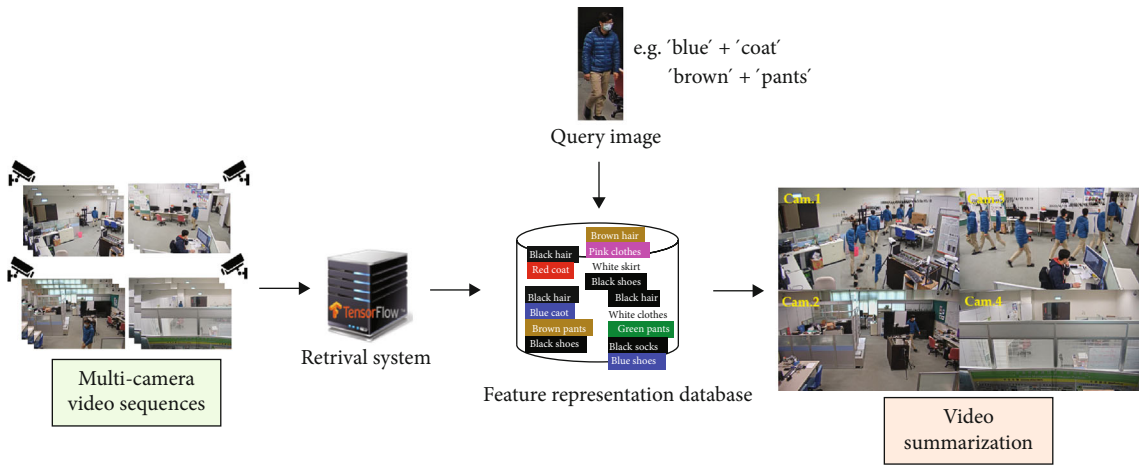


Figure 7: Video summarization process. Video synopsis summarizes video sequences from many cameras.

system. We first describe the dataset and protocol employed to evaluate the system. We also explain the details of the system setup. Finally, we present and analyze the experiment results.

3.1. Implementation Details

3.1.1. Appearance Features with FCN Model. This study used the LIP database to train the FCN model. This database consists of images extracted from the Microsoft COCO database that are classified using 20 labels including the image background. In this paper, we used 10 labels for training: hat, hair, glove, clothes, coat, socks, pants, scarf, skirt, and shoes. There were 29,735 training images. The FCN model parameter settings were as follows: input image size, 550×550 and backend threshold, 0.5. Figure 8 shows the total loss of the FCN model; the training loss curve monotonically decreases with iterations, demonstrating that the model training achieves a stable convergence.

3.1.2. Cloth Color Model. The color model classifies 11 colors: black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow. Our models were trained using the stochastic gradient descent method with 115 examples per

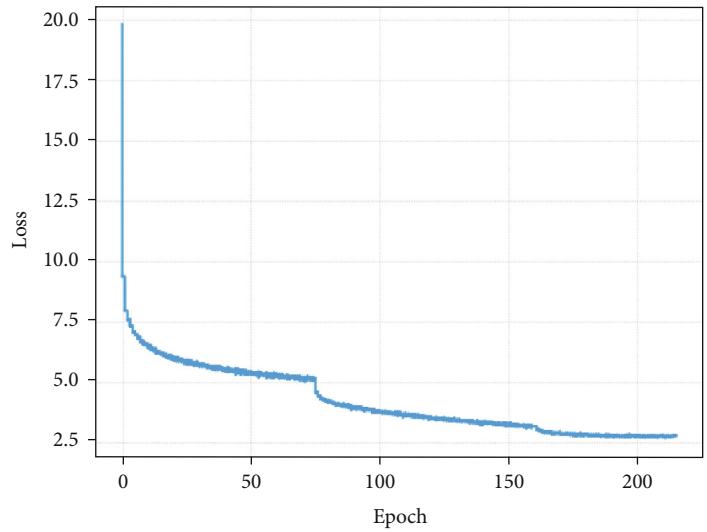


Figure 8: Convergence of the FCN model.

batch, a momentum of 0.9, and a weight decay of 0.0005. A real video dataset was used as training data in the experiments. The clothes of the person in the video were segmented using the FCN model. The images of segmented clothes were resized to $227 \times 227 \times 3$ and labeled to for use as training data. In the training process, data were randomly mirrored to increase the classifier accuracy. We used a learning rate of 0.01 and initialized the weights of the networks using a Gaussian function with $\delta = 0.01$ for connecting weights and a fixed bias value of 0.1. Figure 9 shows the confusion matrix for our clothes color model trained using the RGB color space. The confusion matrix shows that the clothes color model exhibits relatively low accuracy for the blue color class. Some samples of the blue color class were misclassified as the green color class at a rate of approximately 12%. As seen in the dataset, some blue color class samples have more of a blue-green color than a blue color; therefore, the CNN classifier may misclassify these samples as belonging to the green color class. Colors may be misclassified because of the reflection of bright sunlight on clothes or variations in tones or shades of a color resulting, in an appearance similar to another color. Figure 10 shows the convergence of the color model training process, the model accuracies, and the loss curves during training across the number of epochs using segmented images.

3.2. Experimental Analysis. For training the deep learning model and testing, a CUDA-enabled NVIDIA TITAN RTX GPU was used with an Intel Core i9 7900X processor. The model, dataset, and multicamera video surveillance scenes of the proposed system were verified separately. To verify the performance of the pixelwise person segmentation and appearance-based attribute recognition, experiments were conducted on a standard benchmark dataset, namely, DukeMTMC-reID [42], to evaluate the effectiveness of the presented method (see Section 3.2.1). Further, the performance of the proposed system was verified for real multicamera scenarios (see Section 3.2.2).

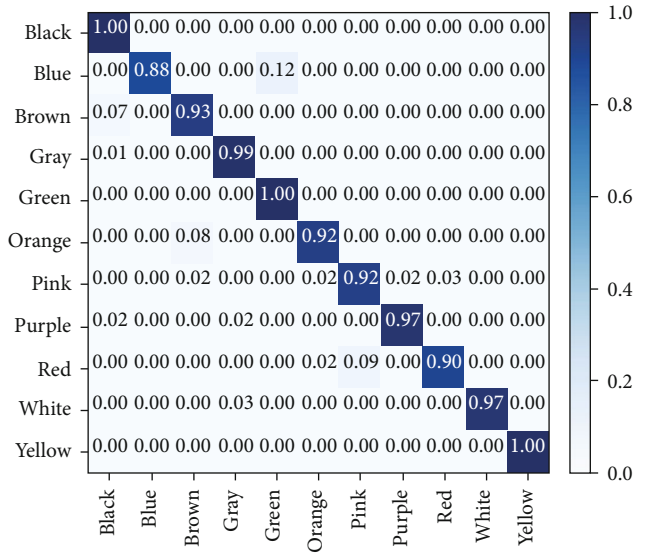


Figure 9: Confusion matrix of our clothes color model. Each cell describes the accuracy (percentage) for each class.

3.2.1. Evaluation of Benchmark Dataset. The DukeMTMC-reID dataset consists of person images extracted from the DukeMTMC [43] tracking dataset. DukeMTMC contains images taken from eight high-resolution cameras, and person-bounding boxes are manually annotated. To illustrate the effectiveness of the proposed method, from the query images of the DukeMTMC-reID dataset, we randomly selected approximately 1,000 images containing 702 different identities as our testing set. For the experiments conducted on this benchmark, manually annotated images were used as the ground truth for testing.

The proposed method was used to calculate the clothing classification rate for the dataset. As shown in Table 1, it achieved an average classification rate of 94.35%. The clothing classification results show that this method provides the highest recognition rates for eight attributes. Further, the

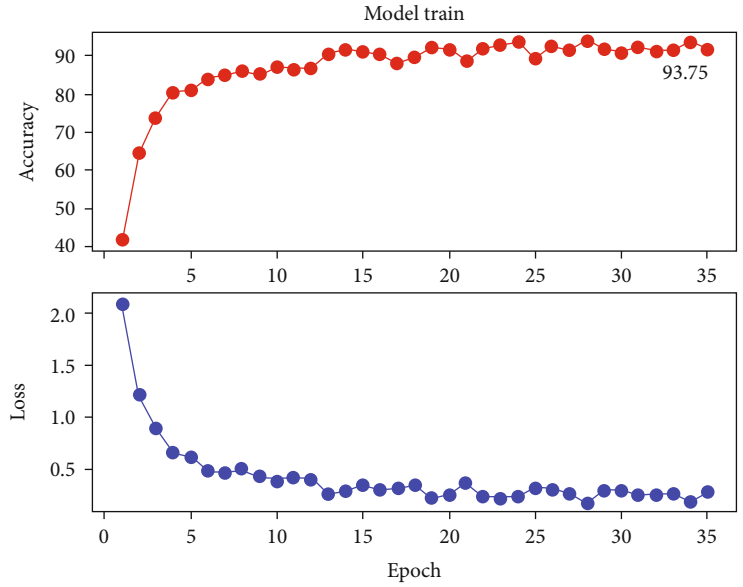


Figure 10: Convergence graphs of the color model on accuracy and loss curves.

recognition rates for ten attributes differ only slightly; for example, the best recognition rate, which is for "skirt," is only 20% higher than the worst recognition rate, which is for "glove." Such results indicate that the appearance-based attribute features obtained using our multi-CNN method are highly uniform and stable.

The proposed method is aimed at identifying appearance attributes or characteristics of the human body, and therefore, high-accuracy segmentation for fitting objects is crucial for recognizing the masked region. Our experiments indicated that the proposed method can effectively localize various parts of human clothing and recognize attribute features even under severe pose variations and occlusions. Figure 11 illustrates the how the proposed method segments the query images from the DukeMTMC-reID benchmark dataset. Figure 9 shows several example images of attribute prediction and recognition results. The multiple attributes of each person are represented by the proposed appearance-based attribute features identified by the multi-CNN method, proving that this method can ably recognize attribute features and process semantic information.

3.2.2. Practical Applications with Multicamera Systems. A multicamera experiment was performed to demonstrate the practicality of the proposed system and to verify its accuracy. Real-time cameras set up in different spaces were used to retrieve person data from various monitoring environment information; this was then input to the system for learning mastery. Figure 12 shows a plan view of the experimental site and the location of the camera rack.

Four cameras were used to perform this experiment. This experiment was conducted to prove that the proposed system can simultaneously identify different person appearance attribute features and analyze spatial and temporal data through pixelwise person segmentation and multi-CNN models in a multicamera environment. The extracted

Table 1: Clothing classification accuracy when applying the system to the DukeMTMC-reID dataset.

Types of clothes	Classification rate (%)
Hat	95.35
Hair	99.90
Glove	80.00
Clothes	98.43
Coat	99.68
Socks	81.82
Pants	99.36
Scarf	90.00
Skirt	100.00
Shoes	98.96
Average	94.35

appearance attributes and color feature representations are imported into the established feature representation database as human association information for the retrieval engine. To fully utilize the information contained by queries in our system, feature representation must have the ability to comprehend time series information such as the location and state change of a person. Figures 13(a) and 13(b) show person segmentation and appearance attribute extraction performed using cameras 1 and 3, respectively.

Figure 14 shows multicamera localization and recognition results obtained using the proposed method on real surveillance images. Each row of the table corresponds to a video from a different camera. The person attribute description is shown in the top-left corner of the surveillance images. These results indicate that the system enables accurate analysis of the appearance attribute features of the same person walking through the field of view of different cameras.

In order to improve readability for the retrieved objects (e.g., key person), the proposed system uses the video

Figure 11: Examples of person attribute recognition. Example images of segmentation masks generated by our FCN model and color model on query images from the DukeMTMC-reID benchmark dataset.

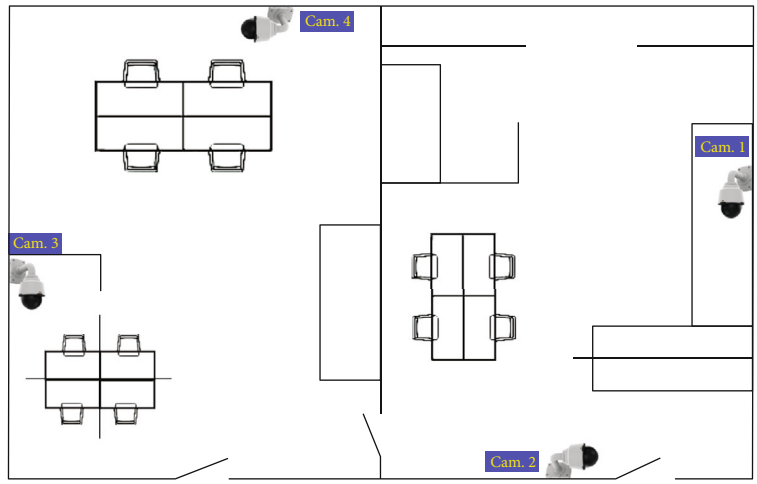
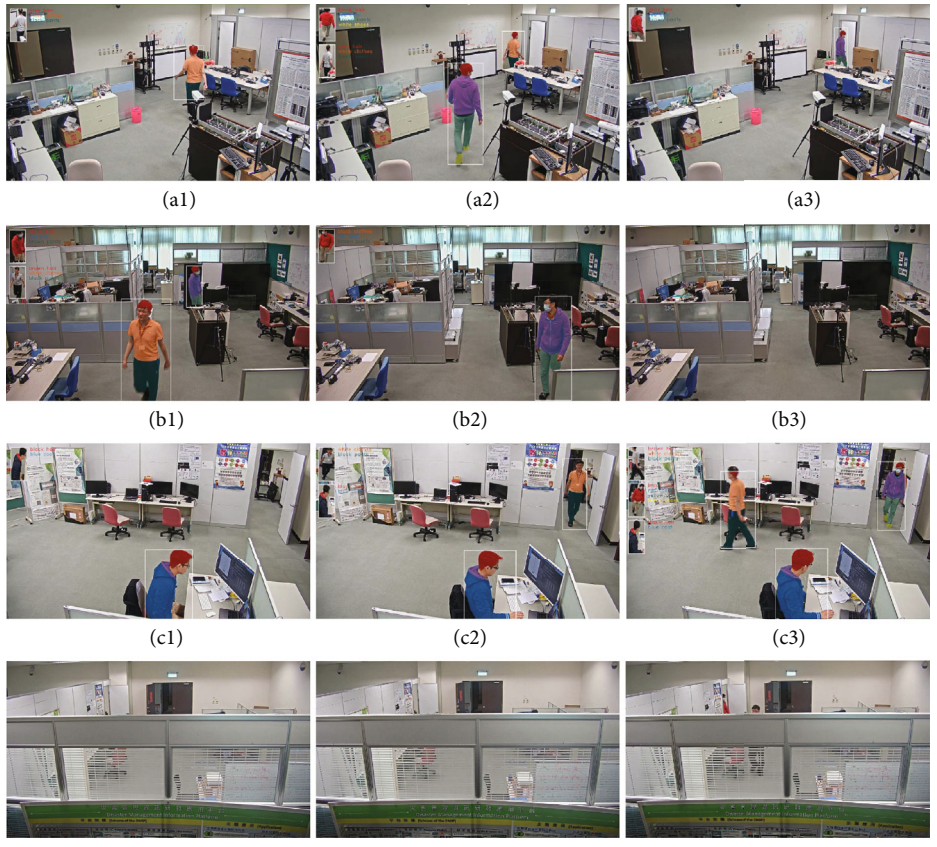


Figure 12: Multiple camera locations and plan view of laboratory room.

summarization method to extract the key frames of the retrieved objects and condenses them in multicamera video surveillance systems. To evaluate the performance of the video summarization method in this paper, we analyzed the results of the two event queries shown in Figure 15: cameras 1–4, "red coat + brown pants" and "white clothes + black

(a) (b)

Figure 13: Results of person segmentation and appearance attribute extraction: (a) camera 1 and (b) camera 3.



(a1) (a2) (a3)

(b1) (b2) (b3)

(c1) (c2) (c3)

(d1) (d2) (d3)

Figure 14: Examples of multicamera localization and recognition results on real surveillance video sequences. Each row represents a camera ID (top to bottom: cameras 1–4). Each column represents a surveillance time series (left to right).

(a) (b)

Figure 15: Results of various event retrievals: (a) cameras 1–4, "red coat + brown pants" and (b) cameras 1–4, "white clothes + block pants".

pants." Furthermore, the proposed summarization method provides a quick overview of the retrieved content in multiple camera surveillance videos. The video summarization method was used to produce dynamic information of a large amount of retrieved objects. Person retrieval was found to accurately present the position of the key person at each moment, and the dynamic behavior of the retrieved person could be visualized effectively.

4. Conclusions

This study proposed a video retrieval system for performing deep learning–based person segmentation and attribute feature representation for person retrieval and summarization from multiple surveillance cameras. The proposed system for person retrieval consists of pixelwise person segmentation, appearance-based multi-CNN, and a video summarization method.

In the person analysis process, we extract the attribute of a masked image from the instance segmentation module for each object of interest. The cluttered background and appearance variation problems that occur when using multiple cameras are solved. Additionally, the proposed person retrieval and summarization method simultaneously retrieves and condenses the key person in video sequences to reduce the need to search through a large number of videos. The proposed method can achieve the following result: (1) To solve search for a specific person in multicamera surveillance, we used automatic instance segmentation framework to perform pixelwise person segmentation and developed a multi-CNN classifier to extract the appearance attribute features of a human masked image to represent a specific person. Therefore, the aim of this paper is to improve the accuracy of attributes classification by using the proposed method and achieve effective retrieval performance. (2) To improve the readability of retrieval image, we also aim to provide an overview of a video sequence without requiring a large amount of viewing time for surveillance. The proposed person retrieval and summarization method simultaneously retrieves and condenses the key person in video sequences to reduce the need to search through a large number of videos. This system was tested on the DukeMTMC-reID benchmark datasets and implemented in real multicamera scenarios. Experiment results demonstrated that the proposed system performs well in multicamera retrieval and surveillance applications and provides satisfactory retrieval performance.

Data Availability

The system parameters used to support the findings of this study are included within the article. The experiment data used to support the study is available upon request to the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 761–774, 2006.

[2] K. K. Verma, P. Kumar, and A. Tomar, "Analysis of moving object detection and tracking in video surveillance system," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1758–1762, New Delhi, India, 2015.

[3] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 602–615, 1998.

[4] J. H. Kim, H. Y. Lim, and D. S. Kang, "An implementation of the video retrieval system by video segmentation," in *2008 14th Asia-Pacific Conference on Communications*, pp. 1–5, Akihabara, Japan, 2008.

[5] Y. Yang, B. C. Lovell, and F. Dadgostar, "Content-based video retrieval (CBVR) system for CCTV surveillance videos," in *2009 Digital Image Computing: Techniques and Applications*, pp. 183–187, Melbourne, VIC, Australia, 2009.

[6] Lijie Liu and Guoliang Fan, "Combined key-frame extraction and object-based video segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 869–884, 2005.

[7] C. Huang and H. Wang, "A novel key-frames selection framework for comprehensive video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 577–589, 2019.

[8] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[9] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Smart surveillance based on video summarization," in *2017 IEEE Region 10 Symposium (TENSYMP)*, pp. 1–5, Cochin, India, July 2017.

[10] D. Lee, S. J. Lee, and Yu-Jeong Seo, "Application of recent developments in deep learning to ANN-based automatic berthing systems," *International Journal of Engineering and Technology Innovation*, vol. 10, no. 1, pp. 75–90, 2020.

[11] Z. Ji, W. Zheng, and Y. Pang, "Deep pedestrian attribute recognition based on LSTM," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 151–155, Beijing, China, September 2017.

[12] J. Xu and H. Yang, "Identification of pedestrian attributes based on video sequence," in *2018 IEEE International Conference on Advanced Manufacturing (ICAM)*, pp. 467–470, Yunlin, Taiwan, November 2018.

[13] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 111–115, Kuala Lumpur, Malaysia, November 2015.

[14] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[15] T. Yang, F. Chen, D. Kimber, and J. Vaughan, "Robust people detection and tracking in a multi-camera indoor visual surveillance system," in *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 675–678, Beijing, China, July 2007.

[16] D. Gerónimo and H. Kjellström, "Unsupervised surveillance video retrieval based on human action and appearance," in *2014 22nd International Conference on Pattern Recognition*, pp. 4630–4635, Stockholm, Sweden, August 2014.

[17] R. Satta, *Dissimilarity-Based People Re-identification and Search for Intelligent Video Surveillance*, Phd thesis, University of Cagliari, Cagliari, Italy, 2013.

[18] N. Zhang, M. Paluri, M. A. Ranzato, T. Darrell, and L. Bourdev, "Panda: pose aligned networks for deep attribute modeling," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637–1644, Columbus, OH, USA, June 2014.

[19] B. Hu, B. Gao, W. L. Woo et al., "A lightweight spatial and temporal multi-feature fusion network for defect detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 472–486, 2021.

[20] L. Ruan, B. Gao, S. Wu, and W. L. Woo, "DeftectNet: joint loss structured deep adversarial network for thermography defect detecting system," *Neurocomputing*, vol. 417, pp. 441–457, 2020.

[21] B. H. D. Koh, C. L. P. Lim, H. Rahimi, W. L. Woo, and B. Gao, "Deep temporal convolution network for time series classification," *Sensors*, vol. 21, no. 2, p. 603, 2021.

[22] H. Galiyawala, M. S. Raval, and S. Dave, "Visual appearance based person retrieval in unconstrained environment videos," *Image and Vision Computing*, vol. 92, p. 103816, 2019.

[23] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[24] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++: better real-time instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2020.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, Venice, Italy, October 2017.

[26] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, Salt Lake City, UT, USA, June 2018.

[27] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6409–6418, Long Beach, CA, USA, June 2019.

[28] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2359–2367, Honolulu, HI, USA, July 2017.

[29] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," in *European Conference on Computer Vision*, pp. 534–549, Amsterdam, The Netherlands, 2016.

[30] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "Masklab: instance segmentation by refining object detection with semantic and direction features," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4013–4022, Salt Lake City, UT, USA, June 2018.

[31] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother, "InstanceCut: from edges to instances with multicut," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5008–5017, Honolulu, HI, USA, July 2017.

[32] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5221–5229, Honolulu, HI, USA, July 2017.

[33] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan, "Proposal-free network for instance-level object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2978–2991, 2017.

[34] A. Newell, Z. Huang, and J. Deng, "Associative embedding: end-to-end learning for joint detection and grouping," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 2277–2287, Long Beach, CA, USA, 2017.

[35] A. W. Harley, K. G. Derpanis, and I. Kokkinos, "Segmentation-aware convolutional networks using local attention masks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5038–5047, Venice, Italy, October 2017.

[36] B. D. Brabandere, D. Neven, and L. V. Gool, "Semantic instance segmentation with a discriminative loss function," https://arxiv.org/abs/1708.02551.

[37] A. Fathi, Z. Wojna, V. Rathod et al., "Semantic instance segmentation via deep metric learning," 2017, https://arxiv.org/abs/1703.10277.

[38] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: speeded up robust features," in *European conference on Computer Vision (ECCV)*, pp. 404–417, Graz, Austria, 2006.

[39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[40] R. F. Rachmadi and I. Purnama, "Vehicle color recognition using convolutional neural network," 2015, https://arxiv.org/abs/1510.07391.

[41] X. Liang, C. Xu, X. Shen et al., "Human parsing with contextualized convolutional neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 115–127, 2016.

[42] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3754–3762, Venice, Italy, October 2017.

[43] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *European Conference on Computer Vision*, 2016, https://arxiv.org/abs/1609.01775.