Pedestrian Tracking based on Improved YOLOv5

Yong Sun, Jiadun Qi, Ruohan Wen, Hao Tian, Cheng Huang and Kao-kin Zao
Department of Information Engineering
The Chinese University of Hong Kong
Hong Kong, China
johnkzao@cuhk.edu.hk

***Abstract*—Object tracking is currently used in many popular scenarios. At this stage, due to the impact of the new crown epidemic, the government wants to deploy an application that can be used for population tracking and flow counting in densely populated places. Based on the current situation, we combined the YOLOv5 model and made some improvements, so that it can be better applied to target tracking scenarios. We add an attention mechanism to the feature extraction network and replaced the NMS in the YOLOv5 algorithm with Adaptive-NMS. The experimental results verify the accuracy improvement of our improvement on pedestrian tracking effect and the further improvement of YOLOv5 speed. Meanwile, the module is flexible, and can be easily applied to other similar scenario tracking models.**

***Index Terms*—Object Tracking, Adaptive-NMS, YOLOv5, Activation Function**

I. Introduction

Object tracking is the automatic determining of position and size information of the target object in a video along the frames. Object tracking methods generally use a smaller-sized search area, rather than the entire frame view, to extract the position and size information of any frame target object. The target tracking is divided into single target tracking and multi-target tracking. Multi-target detection mainly uses the tracking by detection framework. The mainstream approach is to perform target detection for each frame, then perform motion filtering and Reid feature matching, and then use bipartite graph matching to associate before and after frames. The mainstream method of single-target tracking is template matching. The framework used is to train a template-matching siamese network, so that the target template image and subsequent frame matching are associated. Because there is only one target, there is no need for bipartite graph matching, which can be optimized end-to-end. Among them, target tracking involves many fields, such as pedestrian tracking [1][2], drone tracking [3][4], athlete tracking [5][6], etc. There are also many algorithms involved, such as Faster R-CNN [7]. Faster R-CNN does a good job in real-time compared to the previous R-CNN, and some models borrow from Faster R-CNN to achieve better results, such as YOLO [8]. Among these well-known target detection algorithms, one of them is particularly important for their target detection screening, that is, the NMS algorithm [9].

While these classic algorithms have shown great success in deep learning, new attention mechanism algorithms [10] have also been very popular in recent years. Attention mechanism is a resource allocation scheme that allocates computing resources to more important tasks and solves the problem of information overload in the case of limited computing power. In neural network learning, generally speaking, the more parameters of the model, the stronger the expression ability of the model, and the greater the amount of information stored in the model, but this will bring about the problem of information overload. Then, by introducing an attention mechanism, focusing on the information that is more critical to the current task among the many input information, reducing the attention to other information, and even filtering out irrelevant information, the problem of information overload can be solved and the task processing efficiency can be improved. Efficiency and accuracy. This is similar to the human visual attention mechanism. By scanning the global image, the target area that needs to be focused on is obtained, and then more attention resources are devoted to this area to obtain more detailed information related to the target, while ignoring other irrelevant information. Through this mechanism, high-value information can be quickly screened out from a large amount of information using limited attention resources.

In this paper, we combine the adaptive-NMS and attention mechanism network based on the object detection model YOLOv5, and apply the newly proposed method to our project, pedestrian tracking. Experimental results, we compare it with the previous Faster The comparison between R-CNN and YOLOv5, also confirms the efficiency of our method. We will continue to conduct more in-depth research in this area in the future.

Yong Sun is the first author. He is currently pursuing a master's degree in Information Engineering at the Chinese University of Hong Kong.His E-mail is 1155169190@link.cuhk.edu.hk.

Jiadun Qi, Ruohan Wen, Hao Tian and Cheng Huang are currently pursuing a master's degree at the Department of Information Engineering, Chinese University of Hong Kong. Their E-mails are respectively {1155161048, 1155166052, 1155169008, 1155152552} @link.cuhk.edu.hk.

Professor John (Kao-kin Zao) is currently a Professor of Practice in Information Engineering at the School of Engineering, Chinese University of Hong Kong, and the corresponding author of this paper. His research interests are Internet Engineering (IoT & Edge Computing); Wireless Communications and Networking.

II. Methodology

A. YOLOv5

YOLO is the first proposed single-stage object detection algorithm, also known as YOLOv1. The biggest advantage of YOLOv1 is its speed, and its main contribution is to detect the entire image and camera input in real time. YOLOv1 has two disadvantages: one is inaccurate localization, and the other is low recall compared to methods based on candidate bounding boxes.

YOLOv2 solves these problems of YOLOv1. YOLOv2 [11] makes improvements in three aspects: prediction accuracy, detection speed, and the number of recognized objects while continuing to maintain the detection speed. After that, YOLOv3 [12], v4 [13] and the current v5 version [14] came out one after another. They are all based on the previous generation and improved the previous defects. Among them, YOLOv5 has two significant advantages over the previous versions: it is very small and it is quite fast.

The prior detection system of YOLOv3 reuses the classifier or localizer to perform the detection task. Furthermore, the model can be applied to multiple locations and scales of the image. And those regions with higher scores can be regarded as detection results. In addition, compared to other object detection methods, YOLOv3 applies a single neural network to the entire image, which divides the image into different regions and thus predicts the bounding box and probability of each region, which will pass the predicted probability. weighted. My model has some advantages over classifier-based systems. It looks at the entire image when testing, so its predictions take advantage of global information in the image. Unlike R-CNN, which requires thousands of images of a single target, it makes predictions with a single network evaluation. This makes YOLOv3 very fast, typically 1000 times faster than R-CNN and 100 times faster than Fast R-CNN.

Compared with the previous version, YOLOv4 continuously optimizes and adjusts the parameters of the YOLOv3 version, so that each item reaches the optimal solution at that time.

YOLOv5 is a single-stage target detection algorithm. The algorithm adds some new improvement ideas on the basis of YOLOv4, so that its speed and accuracy have been greatly improved. The improvements of v5 at the input end include Mosaic data enhancement, adaptive anchor box calculation, and adaptive image scaling; the improvements in the benchmark network include the fusion of other detection algorithms, such as the Focus structure and the CSP structure; on the Neck network, the v5 target detection network is in Some layers are often inserted between BackBone and the last Head output layer. The FPN+PAN structure is added to Yolov5; on the last Head output layer, its anchor frame mechanism is the same as that of YOLOv4, and the main improvement is the loss function GIOU_Loss during training .

B. Adaptive-NMS

NMS (non maximum suppression) [9] is to suppress elements that are not maximum values and search for local maximum values. Object detection needs to locate the bounding box of the object, and also identify the object in the bounding box. Some candidate boxes will inevitably overlap, and their intersection and union ratio IoU needs to be calculated at this time. The formula of IoU can be expressed as follows:

$$IoU = \frac{A \cap B}{A \cup B}$$

where A and B are two sets, which in the image are the areas of the rectangular candidate boxes of the two candidate boxes. NMS relies on the classifier to obtain multiple candidate boxes, and the probability value of the candidate box belonging to the category, and sorts according to the category classification probability obtained by the classifier. After that, NMS sorts the scores of all boxes, selects the highest score and its corresponding box, and then traverses the remaining boxes. If the overlap area (IOU) with the current highest score box is greater than a certain threshold, NMS deletes the box. Finally, continue to choose the one with the highest score from the unprocessed box and repeat the above process.

But there is a problem: in some cases, a candidate box with a larger overlap with the current highest scoring box is more likely to be a redundant box. Traditional NMS may bring bad results if there is severe occlusion between objects. When the distribution of objects is sparse, NMS can choose a small threshold to eliminate more redundant boxes; when the distribution of objects is dense, NMS chooses a large threshold to obtain higher recall. In this case, we need NMS to have a density prediction module to learn the density of a box. On this basis, the Adaptive-NMS [15] was born. Among them, the algorithm flow chart of Adaptive-NMS is as follows in Fig. 1 [15]:

Input : $\mathcal{B} = \{b_1, .., b_N\}, \mathcal{S} = \{s_1, .., s_N\}$,
$\mathcal{D} = \{d_1, .., d_N\}, N_t$
$\mathcal{B}$ is the list of initial detection boxes
$\mathcal{S}$ contains corresponding detection scores
$\mathcal{D}$ contains corresponding detection densities
N_t is the NMS threshold

begin
  $\mathcal{F} \leftarrow \{\}$
  while $\mathcal{B} \neq$ *empty* **do**
    $m \leftarrow$ argmax $\mathcal{S}$
    $\mathcal{M} \leftarrow b_m$
    $N_{\mathcal{M}} \leftarrow \max(N_t, d_m)$
    $\mathcal{F} \leftarrow \mathcal{F} \bigcup \mathcal{M}; \mathcal{B} \leftarrow \mathcal{B} - \mathcal{M}$
    for b_i *in* $\mathcal{B}$ **do**
      if $iou(\mathcal{M}, b_i) \geq N_t$ **then**
        $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{S} \leftarrow \mathcal{S} - s_i;$
      end Greedy-NMS
      if $iou(\mathcal{M}, b_i) \geq N_{\mathcal{M}}$ **then**
        $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{S} \leftarrow \mathcal{S} - s_i;$
      end Adaptive-NMS
    end
  end
  return $\mathcal{F}, \mathcal{S}$
end

Fig. 1. the algorithm flow chart of Adaptive-NMS.

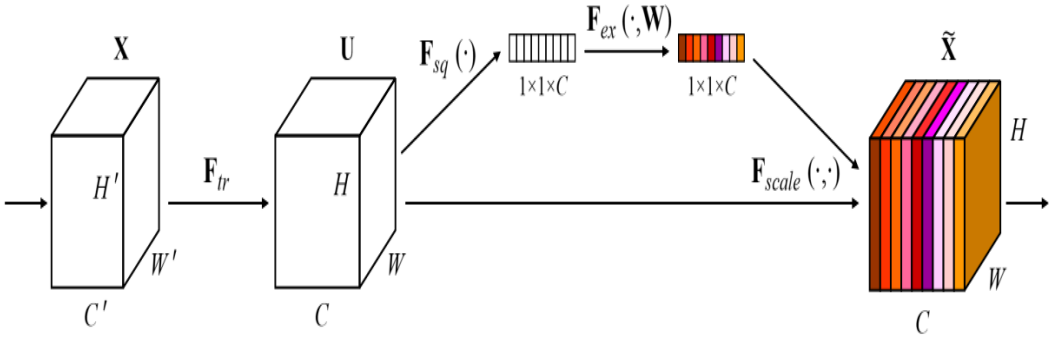

Fig. 2. The structure of Squeeze-and-Excitation block.

Every time this process is performed, the threshold of the NMS will be refreshed again. The continuously adjusted threshold finally enables Adaptive-NMS to better cope with occluded targets. At the same time, this point can also make adaptive-NMS mixed with other types of NMS [16][17][18].

C. Squeeze-and-Excitation block

The motivation of SENet is very simple. The general method is to transmit the weights such as the Feature Map of the network to the next layer. The core idea of SENet is to model the interdependence between channels, and adaptively re-correct the channels through the global loss function of the network. between the characteristic response strengths. SENet consists of a series of SE blocks. The process of a SE block is divided into two steps: Squeeze and Excitation. Among them, Squeeze obtains the global compressed feature vector of the current Feature Map by performing Global Average Pooling on the Feature Map layer, and Excitation obtains the weight of each channel in the Feature Map through two layers of full connection, and uses the weighted Feature Map as the next layer. The input to the network, also known as the SE channel attention mechanism. Since the SE block only depends on the current set of Feature Maps, it can be easily embedded into almost all current convolutional networks [19].

The structure of an SE Block is shown in Fig. 2 [19].

The left half of the network is a traditional convolution transformation, where $\mathrm{F_{tr}}$ is the convolution operation, and ignoring this part will not affect our SENet understanding. U is the Feature Map with the size of X obtained after the $\mathrm{F_{tr}}$ convolution operation, $W \times H \times C$ is the size of the Feature Map, and (W, H) is the number of channels.

AfterSqueeze operation of $\mathrm{F_{sq}}(\cdot)$, the image becomes a $1 \times 1 \times w$ eigenvector, and the value of the eigenvector is determined by U. After the $\mathrm{F_{ex}}(\cdot, W)$, the dimension of the feature vector does not change, but the vector value becomes a new value. These values will be weighted by $\mathrm{F_{scale}}(\cdot, \cdot)$ of U, and the dimensions of $\tilde{X}$ and U are the same.

The role of the Squeeze part is to obtain the global information embedding (feature vector) for each channel of the Feature Map [formula]. In the SE block, this step is achieved by Global Average Pooling (GAP), that is, by averaging the Feature Map of each channel $C, c \in \{1, 2, ...C\}$.

The role of the excitation part is to learn the feature weights of each channel in the $\mathrm{Z_c}$ through the C. Based on the above structure, SE blocks use a gate mechanism composed of two layers of full connection. Therefore, SE blocks can be understood from two perspectives. One is that SE blocks learn the dynamic prior of each Feature Map; the other is that SE blocks can be regarded as Attention in the direction of Feature Map, because the essence of the attention mechanism is to learn a set of weights. value.

D. Our method

Our approach is based on the three algorithms mentioned above, combining them. For Squeeze-and-Excitation block, we replace the Concat module of YOLOv5 with the F-Concat module, the model learns the importance of features from different inputs, concentrating on the important features and ignoring the less important ones. For Adaptive-NMS, We replaced the NMS on the head of YOLOv5 with Adaptive-NMS.

III. Dataset Pre-process and Hardware Equipment

A. Dataset Preprocess

For the dataset, we collected some images from our students themselves and some local passers-by in Hong Kong with their permission. Among them, 350 images were selected as the training set and 150 images were used as the test images. The pictures for the test also include pictures of our students themselves or taken.

The production process of the data set is to input these pictures into Labelme software for labeling to generate jsn files, and then convert them into xml files for training and testing.

B. Hardware Equipment

On the hardware equipment, we use the server provided by the Department of Information Engineering, School of Engineering, Chinese University of Hong Kong. Some parameters

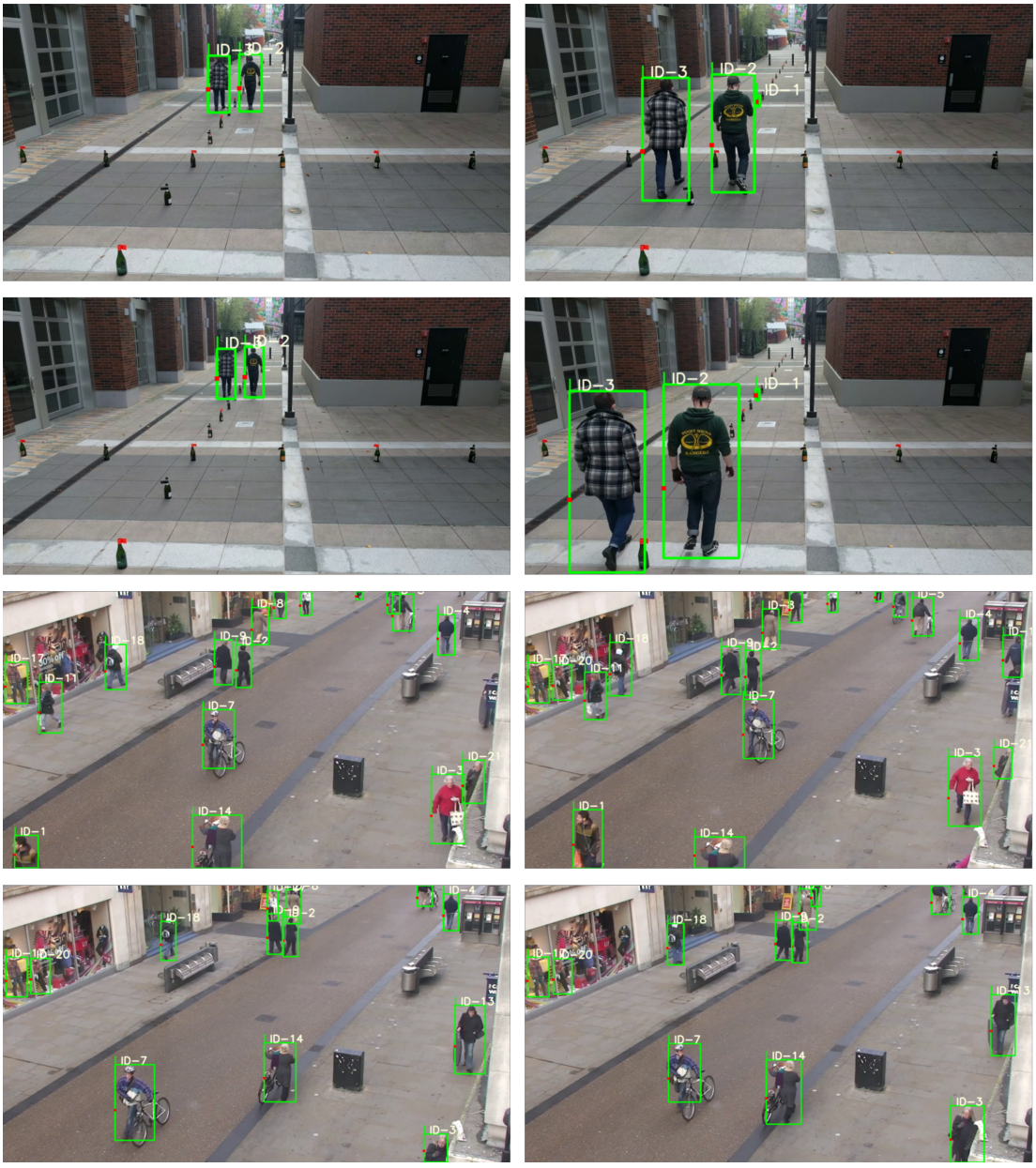

Fig. 3. Experimental results of our method.

of its server are as follows: Ubuntu18.04, 16GB memory, and 2080Ti graphics card.

IV. Experimental Results

As shown in Fig. 3, our method can detect sparse crowd situations as well as general street neighborhoods. Our method can detect tasks and track them even if there are some pushing their own cars or overlapping items in the store. We analyze the results of people who can be coincident with products but detected, which is based on the candidate box density adjustment algorithm of Adaptive-NMS. It makes the confidence score of the place where the character's score is high and the overlapping area is larger continuously reduce, iterate, and finally filter, which can well avoid mistaken deletion. The attention mechanism can better help YOLOv5 track the characteristics of moving targets, thereby helping us to more accurately lock and track the target, even if there are obstacles around the vehicle.

As shown in Table I, our method is compared with other different methods, and the correctness of our method is also verified.

V. Conclusion

In this paper, based on the YOLOv5 model, we improve and replace its activation function and NMS, and use the improved model for pedestrian tracking experiments. And on this basis, other objects have also been extended and tested. Compared with the prototype YOLOv5 and other models, the

TABLE I
THE ACCURACY OF OUR OBJECT DETECTION

Method	NMS_category	Precision	Recall	Accuracy
YOLOv5	NMS	94.1%	90.8%	92.3%
	Softer-NMS	96.3%	92.6%	95.1%
	Adaptive-NMS	96.7%	92.8%	95.4%
Faster R-CNN	NMS	92.1%	89.6%	91.6%
	Softer-NMS	93.2%	90.1%	91.8%
	Adaptive-NMS	92.7%	89.8%	90.6%
Our method	-	97.1%	93.2%	95.9%

[a] The original YOLOv5 and Faster R-CNN did not add an attention mechanism.

experimental results have a good improvement in accuracy and real-time performance. The YOLOv5 model itself is not large, it is extremely easy to be packaged and deployed by mirror clones, and it can also be well trained for other scenarios.

ACKNOWLEDGMENT

Thank you very much for your wonderful cooperation in this class of IEMS 5709. We not only completed intra-group cooperation, but also achieved good cross-group cooperation between groups. We also congratulate us on the success of this project!

REFERENCES

[1] R. Xu and Q. Liu, "Multi-pedestrian tracking for far-infrared pedestrian detection on-board using particle filter," 2015 IEEE International Conference on Imaging Systems and Techniques (IST), 2015, pp. 1-5.

[2] C. He, X. Zhang, Z. Miao and T. Sun, "Intelligent vehicle pedestrian tracking based on YOLOv3 and DASiamRPN," 2021 40th Chinese Control Conference (CCC), 2021, pp. 4181-4186.

[3] J. Park, D. H. Kim, Y. S. Shin and S. Lee, "A comparison of convolutional object detectors for real-time drone tracking using a PTZ camera," 2017 17th International Conference on Control, Automation and Systems (ICCAS), 2017, pp. 696-699.

[4] X. Sun and W. Zhang, "Implementation of Target Tracking System Based on Small Drone," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2019, pp. 1863-1866.

[5] Y. Zhang, Z. Chen and B. Wei, "A Sport Athlete Object Tracking Based on Deep Sort and Yolo V4 in Case of Camera Movement," 2020 IEEE 6th International Conference on Computer and Communications (ICCC), 2020, pp. 1312-1316.

[6] X. Tuo and H. Xie, "Effectiveness of Acute: Chronic Workload Ratio and Oslo Sports Trauma Research Center Questionnaire on Health Problems in Monitoring Sports Load and Injury of Track and Field Athletes," 2021 International Conference on Information Technology and Contemporary Sports (TCS), 2021, pp. 520-523.

[7] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[8] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[9] A Neubeck and LV Gool, "Efficient Non-Maximum Suppression[C]", IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[10] S. Liu and H. Ma, "Combined attention mechanism and CenterNet pedestrian detection algorithm," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 1978-1982.

[11] J. Redmon and A Farhadi, "YOL09000: better faster stronger[C]", Proceedings of the IEEE conference on computer vision and pattern recognition., pp. 7263-7271, 2017.

[12] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement", Computer Science, 2018.

[13] Alexey Bochkovskiy, Chien-Yao Wang and Hong-Yuan Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection", 2020.

[14] T. F. Dima and M. E. Ahmed, "Using YOLOv5 Algorithm to Detect and Recognize American Sign Language," 2021 International Conference on Information Technology (ICIT), 2021, pp. 603-607.

[15] S. Liu, D. Huang and Y. Wang, "Adaptive NMS: Refining Pedestrian Detection in a Crowd," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6452-6461.

[16] N. Bodla, B. Singh, R. Chellappa and L. S. Davis, "Soft-NMS — Improving Object Detection with One Line of Code," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5562-5570.

[17] Y. He, C. Zhu, J. Wang, M. Savvides and X. Zhang, "Bounding Box Regression with Uncertainty for Accurate Object Detection", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2883-2892, 2019.

[18] A. Kumar, G. Brazil and X. Liu, "GrooMeD-NMS: Grouped Mathematically Differentiable NMS for Monocular 3D Object Detection," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8969-8979.

[19] J Hu, L Shen and G. Sun, "Squeeze-and-excitation networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141, 19–21 June 2018.