An Efficient Tracker for Thyroid Nodule Detection and Tracking during Ultrasound Screening

Ting Liu, Xing An, Bin Lin, Yanbo Liu, Wenlong Xu, Yuxi Liu, Longfei Cong, and Lei Zhu(✉)

Shenzhen Mindray BioMedical Electronics, Co., Ltd., Shenzhen, China
zhulei@mindray.com

Abstract. Thyroid tumor is a common disease in clinic. Junior doctors could easily miss or get false detection due to the unclear boundary and similarity between nodules and tissues during thyroid screening. In this paper, we propose an efficient tracker for simultaneously detecting and tracking nodules to assist doctors in examination and improve their work efficiency. An attention based fusion block which adaptively combines the features of previous and current frames is introduced to acquire better detection and tracking result. To increase the detection accuracy, we propose an advanced post-processing strategy instead of using general post-processing methods to train the network to obtain the best prediction. Moreover, a minibatch self-supervised learning module is embedded to reduce the false positive rate (FPR) by strengthening the ability of distinguishing nodules from similar tissues. The proposed framework is validated on a dataset of 1555 thyroid ultrasound movies with 13314 frames. The result of 91% recall with 3.8% FPR running at 30 fps demonstrates the effectiveness of our method.

Keywords: Detection, Tracking, Thyroid ultrasound image, Self-supervised learning

1 Introduction

Thyroid nodules are very common in clinic with the incidence rising rapidly throughout the world. In 2020, 586,202 patients suffered from thyroid cancer, accounting for 2.9% of all cancers [1]. With the growth of health awareness and the widely use of advanced ultrasound equipment, the spotting of thyroid nodules has increased, which brings a great challenge for doctors. Moreover, reviewing large amounts of low-resolution videos is time-consuming, radiologists could lose their concentration which may impact the objectivity on diagnosis. Computer-aided diagnosis can alleviate the workload of doctors and improve the efficiency of their work [2]. Therefore, the development of an automatic and accurate analysis method is necessary for thyroid ultrasound screening.

In recent years, with the development of deep learning, researchers have been investigating convolutional neural networks on thyroid ultrasound image analysis, such as nodule classification and detection. For the task of thyroid image classification, Chi *et al.* [2] fine-tuned a GoogLeNet to extract features of ROIs and predicted the malignancy using the Cost-Sensitive Random Forest algorithm. Ma *et al.* [3] used two pre-trained convolutional neural networks to fuse low-level and high-level features for the classification of thyroid nodules. For the challenge of nodule detection, Abdolali *et al.* [4] enhanced the reliability of detection on a small dataset by modifying Mask R-CNN architecture and combining it with transfer learning. Li *et al.* [5] developed a detector based on Faster R-CNN, and adopted strategies such as layer concatenation and spatial constraint to reach a higher accuracy. Xie *et al.* [6] proposed an SSD based neural network with redesigned loss function and post-processing method to improve the detection recall rate. Wang *et al.* [7] presented an artificial intelligence diagnosis system based on the YOLOv2 to locate and classify nodules simultaneously. Generally, typical techniques widely used in thyroid nodule detection require post-processing method such as non-maximum suppression (NMS) to obtain the optimal prediction in inference.

Although many researches have been done in thyroid ultrasound, the following problems still remain. (i) The majority of methods only focuses on detection in images, ignoring the relationship between previous and current frames in movies. (ii) The performance of NMS used in most framework is limited due to the fixed rules set in advance. Nested predictions of thyroid nodules can still remain after NMS as shown in the first row of Fig. 1, where the tissue inside or outside a nodule seems like another lesion. (iii) Many normal tissues can be recognized as nodules due to the similarity between them as shown in the second row of Fig. 1, but less attention is paid to differentiate them.

In view of above issues, we propose an efficient tracker for nodule detection and tracking during thyroid ultrasound scanning. Firstly, we introduce an attention based fusion block to adaptively combine the features of previous and current images to get better detection and tracking result. Secondly, an advanced post-processing strategy that trains the model rather than uses NMS method to find the optimal result is proposed to improve detection accuracy. Finally, a minibatch self-supervised learning module is embedded as a branch in training period to enhance the ability of discriminating nodules from similar tissues, thus to reduce the false positive rate (FPR).

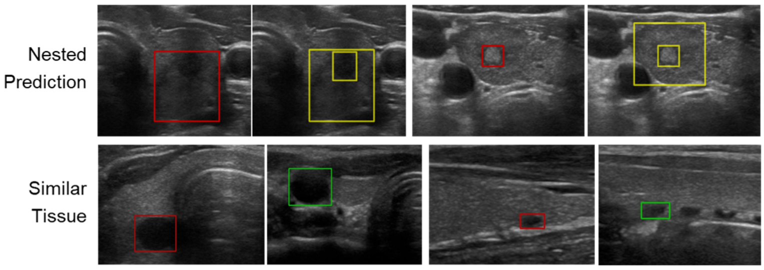

Fig. 1. Illustration of existing problems. Red: the ground truth. Yellow: nested prediction. Green: normal tissues that are similar to nodules.

2 Methodology

2.1 Overall Architecture

The proposed network is primarily based on CenterTrack [8] which sets a new state of the art on both MOT17 and KITTI datasets. Our method is illustrated in Fig. 2. The network takes the current frame, the previous frame and the heatmap [8] generated from objects in the prior frame as input, and outputs the predicted rectangles, the classification probability of each prediction and the center offset of tracked boxes in adjacent frames. The inputs are combined by a fusion module before being fed into the backbone network which is an encoder-decoder structure, and we adopt ResDCN-18 [8] as the backbone in this work. The outputs are divided into two parts, 1) the predicted boxes and classification probabilities to generate the detection results, and 2) the tracking offsets to determine whether the nodules on the prior and current frame are the same one.

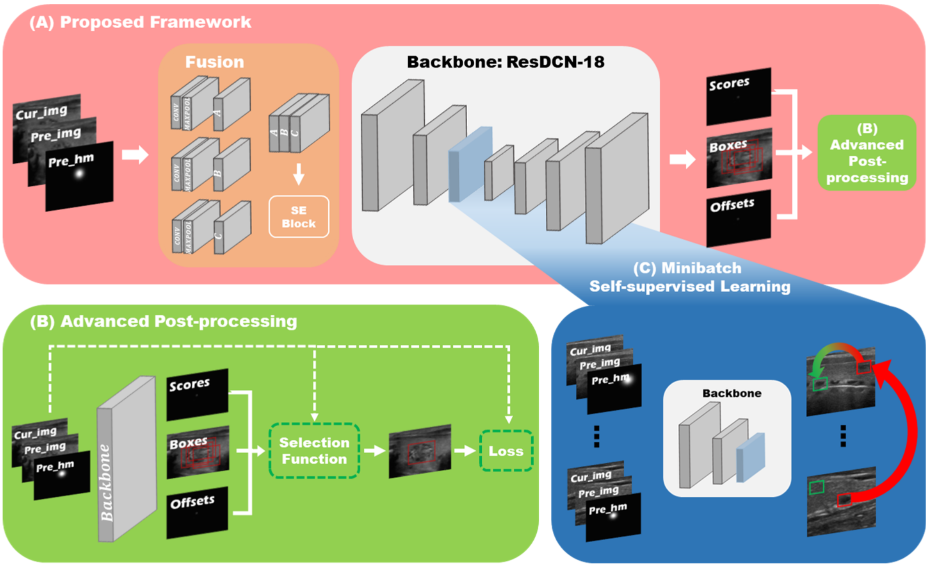

Fig. 2. Illustration of our method. (A) The proposed framework. (B) The advanced post-processing module. (C) The minibatch self-supervised learning module. 'Cur_img', 'Pre_img' and 'Pre_hm' are abbreviations of current image, previous image and heatmap. 'Boxes', 'Scores' and 'Offsets' represent the predicted boxes, the classification probability of each prediction and the offset of the prediction on the current image from the previous frame.

2.2 Fusion Module

The fusion module includes convolution layers and max-pooling layers. It extracts the feature of each input separately and generates three feature maps with size of 1/4 of the input. CenterTrack fuses the feature maps by adding them together and makes each map contribute the same to the next stage. However, we argue that the prior image and heatmap are auxiliary inputs for guiding the model to detect nodules in the current frame. The goal of comprising them is to enhance the performance of detection and tracking. Hence, the feature of current frame should be more important than the other two. Therefore, we concatenate the three feature maps and utilize a squeeze-and-excitation (SE) block to adjust the importance of each channel adaptively. The parameter of reduction in SE block is set to 16 as [9].

2.3 Advanced Post-processing Module

In the detection and tracking task, the optimal prediction should not only consider the classification possibility but also the intersection over union (IoU) and tracking offset with the ground-truth. Inspired by OneNet [10], we introduce an advanced post-processing module to find the optimal result considering of all three aspects above before calculating the loss, thus the model can be trained to acquire the best prediction directly. Meanwhile, no NMS method is required in inference, and end-to-end strategy is achieved. As shown in Fig. 2 (B), the advanced post-processing module takes predictions from the backbone as input. Before calculating the loss, the selection function calculates a score of each prediction, and only the one with smallest score is considered as the correct prediction and the others are assumed as wrong predictions when computing the loss. The selection function is defined as:

$$S = \lambda_{cls} * S_{cls} + \lambda_{L1} * S_{L1} + \lambda_{giou} * S_{giou} + \lambda_{track} * S_{track} \quad (1)$$

where S_{cls} is the focal loss [11] of predicted classifications and ground truth category labels, S_{L1} and S_{giou} are the L1 loss and the GIoU [12] loss between normalized predictions and ground truth boxes, respectively. S_{track} is the L1 loss between predicted tracking offsets and the real displacement of tracked objects. λ_{cls}, λ_{L1}, λ_{giou}, and λ_{track} are coefficients of each component. Following [8,10], λ_{cls}, λ_{L1}, λ_{giou} and λ_{track} are set to 2, 5, 2 and 2, separately.

When calculating detection and tracking loss, each ground truth box only corresponds to one predicted result. The loss function is similar to the selection function, and defined as:

$$L_{dt} = \mu_{cls} * L_{cls} + \mu_{L1} * L_{L1} + \mu_{giou} * L_{giou} + \mu_{track} * L_{track} \quad (2)$$

where L_{cls}, L_{L1}, L_{giou}, and L_{track} have the same definition as S_{cls}, S_{L1}, S_{giou} and S_{track}, separately. And the coefficients μ_{cls}, μ_{L1}, μ_{giou} and μ_{track} are equal to λ_{cls}, λ_{L1}, λ_{giou}, and λ_{track}, separately.

2.4 Minibatch Self-supervised Learning Module

To reduce false positive rate, we embed a minibatch self-supervised learning module as a branch in the training period to differentiate nodules from similar tissues. As shown in Fig. 2 (C), the features of the positive (nodules) and negative (normal tissues) samples are obtained in the down-sampling stage. Inspired by [13], a batch-based similarity loss function is proposed to make the Euclidean distance of positive features closer to the most dissimilar positive features and further away from the most similar negative features in a batch. Besides, a constant is set as threshold to ignore tissues that are dissimilar to nodules in nature. The batch-based similarity loss function is defined as:

$$L_{bs} = \frac{1}{2}\left\|f(p_i), S_p(p_i)\right\|^2 + \frac{1}{2}[max(0, margin - \|f(p_i), S_n(p_i)\|)]^2 \tag{3}$$

where $f()$ denotes the global average pooling. $f(p_i)$ is the feature vector of the i^{th} nodule in a batch. $S_p(x)$ and $S_n(x)$ obtains the most dissimilar positive and the most similar negative feature vector to the $f(x)$ in a batch, separately. To be specific, $S_p(p_i)$ is a positive feature vector whose Euclidean distance is the farthest to $f(p_i)$ in the batch and $S_n(p_i)$ is a negative feature vector that has the nearest Euclidean distance to $f(p_i)$ in the batch. $\|\ \|$ represents the Euclidean distance. Margin sets the distance threshold between $f(p_i)$ and $S_n(p_i)$. The loss function will ignore the samples whose Euclidean distance of $f(p_i)$ and $S_n(p_i)$ is larger than the margin as it indicates that the positive sample and the selected negative sample are not similar. Following [13] margin is set to 10.

Furthermore, we randomly produce false samples on images instead of using manual annotation. For a training batch, if the portion of a nodule on an image is less than α, a negative sample that is not intersected with the nodule will be randomly generated, with the size of β times of the nodule. We set $\alpha = 0.3$ because the ratio of false positives is small and less than 1/3 of an image in our dataset, and $\beta \in [0.9, 1.1]$ to maintain the balance between positive and negative samples.

The total loss is summarized as:

$$\mathcal{L} = L_{dt} + \gamma * L_{bs} \tag{4}$$

where L_{dt} is the detection and tracking loss and L_{bs} denotes the minibatch self-supervised learning loss. γ is a factor and set to 0.2, following [8,14].

3 Experiments

Dataset. We validated the proposed method on 1555 thyroid ultrasound movies from 1555 patients collected via Mindray Resona 7. There are totally 13,314 frames that extracted from the movies at a fixed (3, 4 or 5) interval. All the images were annotated by five doctors firstly and the final annotations were reviewed by an experienced doctor. We calculated the size of nodules (varied from 462 pixels to 232,672 pixels, mean: 29,979 pixels) and classified the movies into 3 categories according to the tri-sectional quantile of size: small (< 1503 pixels), middle (from 1503 pixels to 6923 pixels), and large (> 6923 pixels). We randomly split the movies into 80%, 10%, and 10% for training (1244 movies, 10597 images), validation (155 movies, 1363 images) and testing (156 movies, 1354 images) following the stratified sampling.

Implementation Details. We cropped all data and only the ultrasound image contents were remained. All images were resized to 544 × 640 according to the mean image size of our data. Data augmentations were applied including random horizontal flipping, shifting, scaling, and brightness and contract transformation. Adam optimizer with a learning rate of 5e-5 was utilized for training. We trained the network for 300 epochs with batch size of 32 and chose the model with highest AP50 on validation set to do testing. All experiments were performed on a 12GB NVIDIA TITAN V GPU.

Quantitative and Qualitative Analysis. We measured the FPR (numbers of false positives divided by numbers of images), recall and precision (Prec) to evaluate the detection performance, and recall in tracking (RCLL), mostly tracked (MT) and mostly lost (ML) for tracking performance [15]. The IoU score is set to 0.3 when calculating recall and precision. The novel CenterNet [16], Centertrack, and YoloSiam [17] were re-implemented and evaluated on the same dataset for comparison. Results are shown in Table 1. Our method outperforms all these algorithms not only in above metrics but also in inference speed.

As shown in Fig. 3, our method works well even if the nodule has an undefined margin or is very tiny that less experienced doctors might neglect. Moreover, our method can continuously track on the nodules appear through the whole films. Additionally, our method also works well on multi-target.

Table 1. Performance comparison on the thyroid dataset (%).

Method	FPR	Recall	Prec	AP	AP50	RCLL↑	MT↑	ML↓	Speed
CenterNet	16.1	79.5	83.7	45.2	74.8	NA	NA	NA	28fps
YoloSiam	21.6	76.3	78.2	41.7	70.5	66.8	61.0	25.0	16fps
CenterTrack	13.5	82.1	85.3	48.1	77.4	76.6	65.4	16.0	28fps
Ours	**3.8**	**91.2**	**95.8**	**56.8**	**88.2**	**86.4**	**76.9**	**5.8**	**30fps**

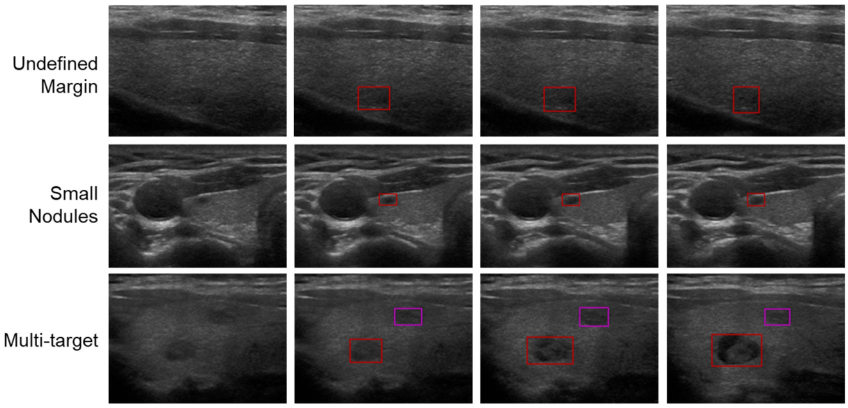

Fig. 3. Detection and tracking results. The first column is an example frame, and the following three columns refer to the second, the forth, and the sixth frame after it.

Ablation Study. An ablation study is conducted on pretrained ResDCN-18 to compare our approach with a representative baseline method. The quantitative results are shown in Table 2. The improvement of the fusion module (F) obtains a better FPR (12.9%) and recall (83.3%) with the same backbone compare to baseline. After using the advanced post-processing module (AP), the FPR is reduced to 8.4% and recall is 2.1% higher than before. The further improvements of FPR from 8.4% to 3.8% and recall from 85.4% to 91.2% indicate the minibatch self-supervised learning module (MSL) has strengthened the ability of distinguishing nodules from similar tissues. We experiment on another backbone (DLA-34) to validate the effectiveness of the proposal. The results also demonstrate the advantage of our method (Table 3).

Table 2. Results of ablation studies based on pretrained ResDCN-18 (%).

Method	FPR	Recall	Prec	AP	AP50	RCLL↑	MT↑	ML↓
Baseline	13.5	82.1	85.3	48.1	77.4	76.6	65.4	16.0
+F	12.9	83.3	85.8	48.3	78.2	78.2	67.9	14.7
+F+AP	8.4	85.4	90.5	51.0	82.5	80.9	71.2	10.3
+F+AP+MSL	**3.8**	**91.2**	**95.8**	**56.8**	**88.2**	**86.4**	**76.9**	**5.8**

Table 3. Results of ablation studies based on DLA-34 (%).

Method	FPR	Recall	Prec	AP	AP50	RCLL↑	MT↑	ML↓
Baseline	14.1	81.3	85.1	46.6	75.7	76.1	63.5	17.9
+F	12.6	82.7	86.0	48.3	78.0	77.8	67.3	15.4
+F+AP	9.8	84.9	88.9	50.7	81.2	79.8	69.2	11.5
+F+AP+MSL	**4.9**	**88.5**	**94.6**	**55.2**	**87.5**	**83.8**	**72.4**	**7.7**

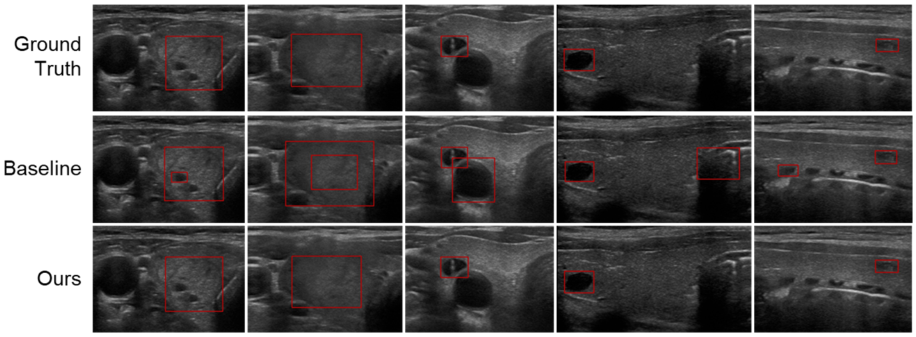

Fig. 4. Results comparison of baseline and ours.

We calculate the mean channel weight of each input given by the SE block in fusion module. The current frame carries the highest weight of 0.52, the previous frame holds the middle (0.49), and the previous heatmap scores the smallest one at 0.46, which validates our hypothesis. Moreover, the first two columns in Fig. 4 illustrate the proposed method can effectively solve the problem of nested predictions that common scheme cannot handle. The last three columns in Fig. 4 present cases that the baseline detects normal tissues as nodules but ours not, which demonstrates our methods can differentiate nodules from the similar tissues better. Meanwhile, Fig. 5 compares the proportion of false positives in testing set, which can show the improvement of our method in FPR clearly.

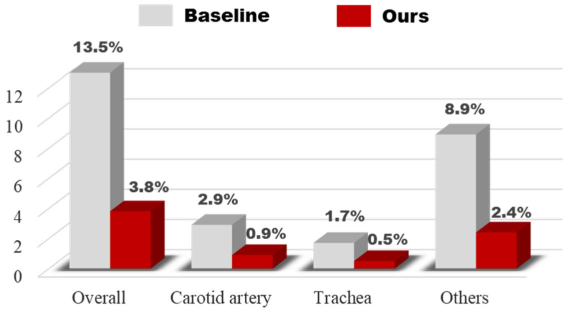

Fig. 5. False positive comparison.

4 Conclusion

To conclude, we propose an efficient tracker to detect and track nodules simultaneously during thyroid ultrasound screening. The attention based fusion block adaptively combines features of the previous and current frames, thus better detection and tracking result is acquired. Moreover, the advanced post-processing mechanism that trains the model instead of using NMS method to select the optimal prediction successfully boosts the detection accuracy. Additionally, the minibatch self-supervised learning module effectively reduces the FPR by enhancing the ability of distinguishing nodules from similar tissues. The result of fast speed, high accuracy, and low FPR obtained from experiments on a challenging and representative dataset reveals a great potential of our system in clinic.

References

1. World Health Organization: Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020. International Agency for Research on Cancer. Geneva: World Health Organization (2020)
2. Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M.: Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. Journal of digital imaging, 30(4), 477-486 (2017)
3. Ma, J., Wu, F., Zhu, J., Xu, D., Kong, D.: A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics, 73, 221-230 (2017)

4. Abdolali, F., Kapur, J., Jaremko, J. L., Noga, M., Hareendranathan, A. R., Punithakumar, K.: Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. Computers in Biology and Medicine, 122, 103871 (2020)
5. Li, H., Weng, J., Shi, Y., Gu, W., Mao, Y., Wang, Y., Zhang, J.: An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. Scientific reports, 8(1), 1-12 (2018)
6. Xie, S., Yu, J., Liu, T., Chang, Q., Niu, L., Sun, W.: Thyroid nodule detection in ultrasound images with convolutional neural networks. In: 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1442-1446 (2019)
7. Wang, L., Yang, S., Yang, S., Zhao, C., Tian, G., Gao, Y., Lu, Y.: Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World journal of surgical oncology, 17(1), 1-9 (2019)
8. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision, pp. 474-490. Springer, Cham (2020)
9. J. Hu, L. Shen and G. Sun.: Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7132-7141 (2018)
10. Sun, P., Jiang, Y., Xie, E., Yuan, Z., Wang, C., Luo, P.: OneNet: Towards End-to-End One-Stage Object Detection. arXiv preprint arXiv:2012.05780 (2020)
11. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980-2988 (2017)
12. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized inter-section over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 658-666 (2019)
13. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815-823 (2015)
14. Tim, S., Ian, G., Wojciech, Z., Vicki, C.: Improved techniques for training GANs. Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), 2234–2242 (2016).
15. Tim, S., Milan., Anton., et al.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
16. Xingyi, Z., Dequan, W., Philipp, K.: Objects as Points. *CVPR* (2019).
17. Labit-Bonis, C., Thomas, J., Lerasle, F., Madrigal, F.: Fast Tracking-by-Detection of Bus Passengers with Siamese CNNs, 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-8 (2019).