Box-Level Tube Tracking and Refinement for Vehicles Anomaly Detection

Jie Wu, Xionghui Wang, Xuefeng Xiao, Yitong Wang
ByteDance Inc.
{wujie.10, wangxionghui.kk, xiaoxuefeng.ailab, wangyitong}@bytedance.com

Abstract

Traffic Anomaly detection is an essential computer vision task and plays a critical role in video structure analysis and urban traffic analysis. In this paper, we propose a box-level tracking and refinement algorithm to identify anomaly detection in road scenes. We first link the detection results to construct candidate spatio-temporal tubes via greedy search. Then the box-level refinement scheme is introduced to employ auxiliary detection cues to promote the abnormal predictions, which consists of spatial fusion, still-thing filter, temporal fusion, and feedforward optimization. Still-thing filter and feedforward optimization employ complementary detection concepts to promote the abnormal predictions, which helps determine an accurate abnormal period. The experimental results show that our approach is superior in the Traffic Anomaly Detection Track test set of the NVIDIA AI CITY 2021 CHALLENGE, which ranked second in this competition, with a 93.18% F1-score and 3.1623 root mean square error. It reveals that the proposed approach contributes to fine-grained anomaly detection in actual traffic accident scenarios and promoting the development of intelligent transportation.

1. Introduction

Traffic Anomaly detection is a fundamental computer vision task and plays a critical role in video structure analysis and potential downstream applications, e.g., accident forecasting, urban traffic analysis and evidence investigation. Traffic anomaly detection has been extensively studied in the computer vision field for a long time [6, 11, 7, 5, 10, 3, 4, 19, 16, 22, 20, 23]. These researches have been conducted to leverage a series of statistic patterns to model abnormal concepts, e.g., Hidden Markov Model [7, 5], Markov Random Field [6, 11], sparse reconstruction [10, 3, 21, 9] and autoencoders [4, 19].

Vehicle anomaly detection is a particular fine-grained traffic anomaly detection, which aims to detect anomalies such as lane violations, wrong-direction driving, and so on. NVIDIA AI CITY CHALLENGE 2018[12], 2019[13], 2020 [14] and 2021 held an vehicle anomaly detection challenge for traffic scenarios. Specifically, each participating team submits detected anomalies, including wrong turns, wrong driving direction, lane change errors, and all other anomalies based on video feeds available from multiple cameras at intersections and along highways. This challenge popularized fine-grained anomaly detection in actual traffic accident scenarios and promoted the development of intelligent transportation.

In NVIDIA AI CITY CHALLENGE 2021, we follow the assumption from [8] that if a stopped vehicle stays longer than the traffic light signal period, it can be regarded as an abnormal event. In this work, we propose a box-level tracking and refinement framework to perform the traffic anomaly detection. The framework contains the extraction of hypothetical differential mask, backward background modeling to eliminate dynamic traffic disturbance, the multi-stage detection model to obtain still vehicles, a box-level tracking mechanism to construct candidate abnormal tubes, and a refinement scheme to promote a more accurate abnormal period. In order to avoid the interference of roadside parking, we use the video motion information to extract and segment the pixel-level hypothetical abnormal mask. We perform backward background modeling based on the Gaussian Mixture Model (GMM) to eliminate moving vehicles, hence the motionless vehicles are easier to detect. Then a multi-stage vehicle detection model is exploited to detect vehicles in the video frames. The box-level tracking branch links the detected boxes and constructs the tube to enclose the trajectory of the anomaly. The refinement scheme is crafted to make up for the performance loss of the detector and alleviate the problem of false detection. To sum up, the main contributions are summarized as follows:

• We present a box-level tracking and refinement framework, which employs different dimension detection results to predict the abnormal period, which helps to take good advantage of comprehensive information for determining an accurate abnormal time.

• The novel box-level refinement scheme is designed to promote the abnormal predictions, which involves spatial

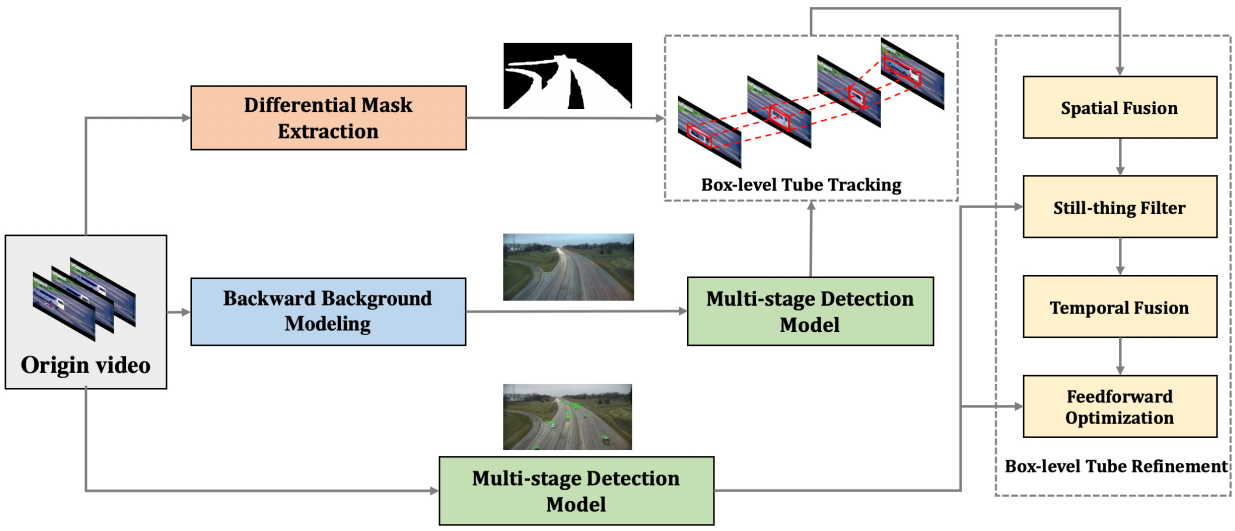

Figure 1. The illustration of box-level tracking and refinement framework. We employ detection results from original video frame and from backward background modeling to obtain accurate abnormal predictions.

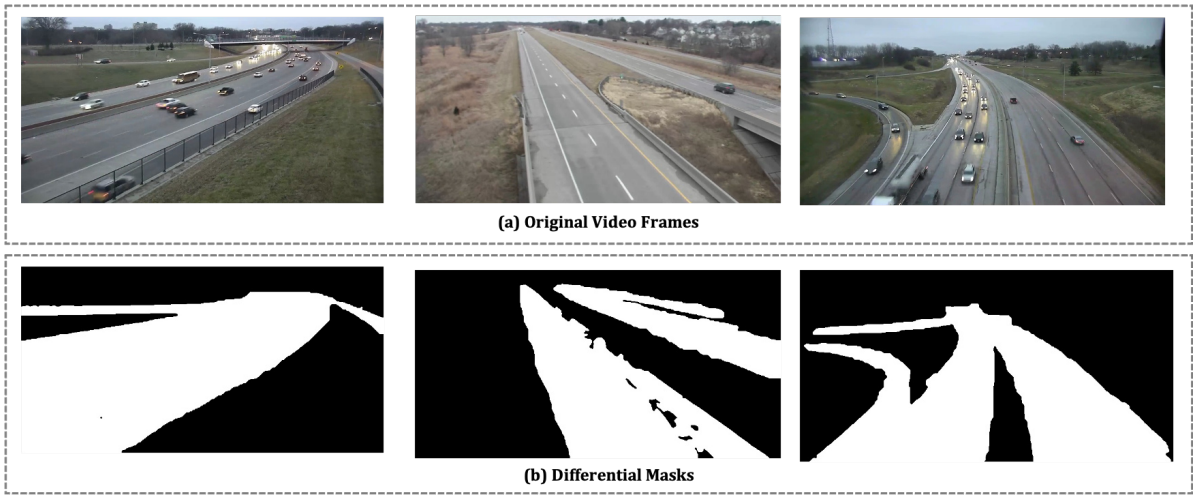

Figure 2. Examples of differential mask. The obtained mask helps to segment the main road precisely.

fusion, still-thing filter, temporal fusion and feedforward optimization. This strategy compensates for the false detection and significantly improves the accuracy of the predictions.

• Based on the above technical points, we evaluated our method on the Traffic Anomaly Detection Track test set of the NVIDIA AI CITY 2021 CHALLENGE, We ranked second among all participating teams, and we obtain the F1-score metric at 0.9318 and the RMSE metric at 3.1623.

2. Methodology

Figure 1 illustrates the proposed framework and its modular components. In the following sections, we first describe the backward background modeling and the extraction of the differential abnormal masks in sections 2.1 and 2.2, respectively. Then the multi-stage detection model is illustrated in section 2.3. Section 2.4 introduces the proposed box-level tracking and refinement approach. Box-level refinement scheme consists of spatial fusion, still-thing filter, temporal fusion and feedforward optimization. Different granularity detection results are employed to obtain and refine the temporal abnormal cues.

2.1. Extraction of Differential Mask

It is possible for the detector to recognize other objects not on the main road as vehicles, resulting in false detection. In order to avoid the interference of roadside parking, we use the video motion information to extract and

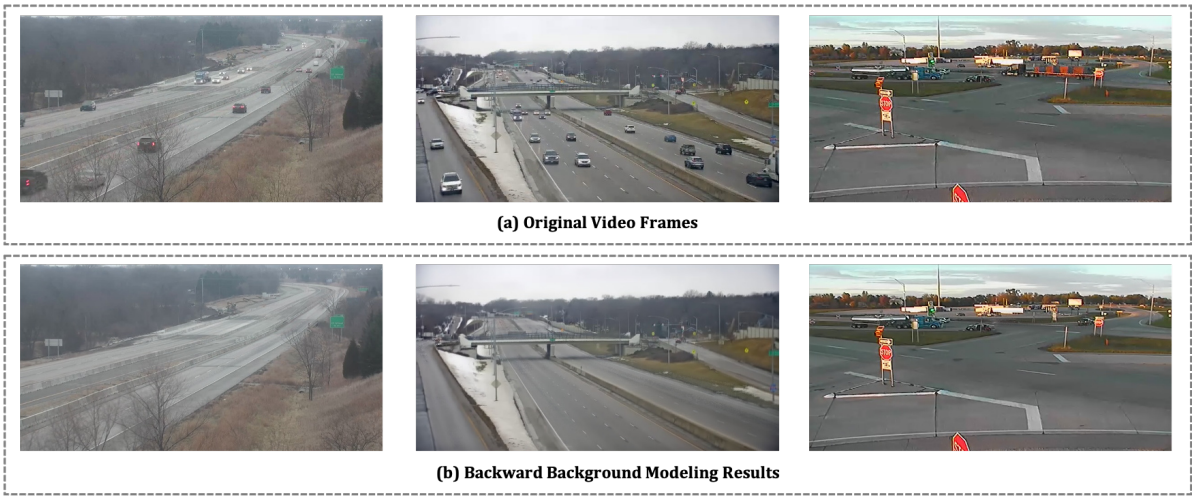

Figure 3. Examples of backward background modeling. From the figure we can see that the background modeling helps to filter the traffic flow and retain the motionless vehicles.

segment the pixel-level hypothetical abnormal mask. We analyze motion differences between two frames to extract the abnormal mask. Specifically, there are k interval frames between these two frames, if the differences exceed η in the pixel-level dimension, we consider that the corresponding area has moving objects and retain this area. Finally, we combine these areas to obtain the differential mask. Figure2 shows some results of the differential mask, we can observe that the obtained mask helps to segment the main road precisely.

2.2. Backward Background Modeling

Whether it is abnormal parking or a traffic accident, it will be accompanied by an abnormal stay of the vehicle. When the vehicle stays for a long time, it will merge into the background information when performing the background modeling, hence we can detect the abnormal vehicle in the background information directly. This paper uses MOG2 [24, 25] for background modeling. MOG2 is an adaptive Gaussian mixture model, which is better than other background extraction algorithms MOG, MOG2, GMG in terms of foreground continuity and computing time.

In this paper, we adopt the background modeling in the backward direction, since we find that the forward background modeling is easy to predict the lagging time in the temporal dimension. As shown in Figure 3, we can observe that the backward background modeling helps to filter the traffic flow and retain the motionless vehicles.

2.3. Vehicle Detection

Object detection is a long-standing topic in the field of computer vision, aiming to detect objects of predefined categories. Recent CNN-based detection methods can be divided into anchor-based and anchor-free detectors. Anchor-based detectors usually fall into one-staged and two-staged methods, while anchor-free detectors consist of keypoint-based and center-based methods. Two-stage methods refine anchors several times more than one-stage methods, hence it has more accurate results while one-stage methods have higher computational efficiency. And state-of-the-art results on common detection benchmarks are still held by anchor-based detectors. Faster-RCNN [15] is a classical two-stage detection framework, which consists of a region proposal network (RPN) and a region-based prediction network(R-CNN). By extending Faster-RCNN, Cascade R-CNN [1], a multi-stage object detection framework was proposed to avoid the problems of overfitting during training and quality mismatch at inference. In the situation of vehicle anomaly detection, recall of vehicles plays a essential role. So we train a Cascade R-CNN detector which adopts ResNeXt [18] with the depth of 101 and groups of 64 as the backbone to extract semantic features. We also employ FPN with layers of 5 to improve detection performance on small objects. A three-stage cascade with different IoU thresholds is adopted to obtain higher quality detection results.

We conduct the experiment on the PyTorch framework. The model is trained on AICITY2021 track4 training videos and performs inference on the origin test videos and videos after backward background modeling, respectively. Detection results of backward background modeling are used to find the abnormal stopped vehicles, while detection results of origin videos are leveraged to filter some still things such as road signs or parked cars. Some examples of detection results are shown in Figure 4.

2.4. Box-level Tube Tracking and Refinement

In this paper, we design a box-level tube tracking and refinement algorithm to analyze the candidate abnormal ve-

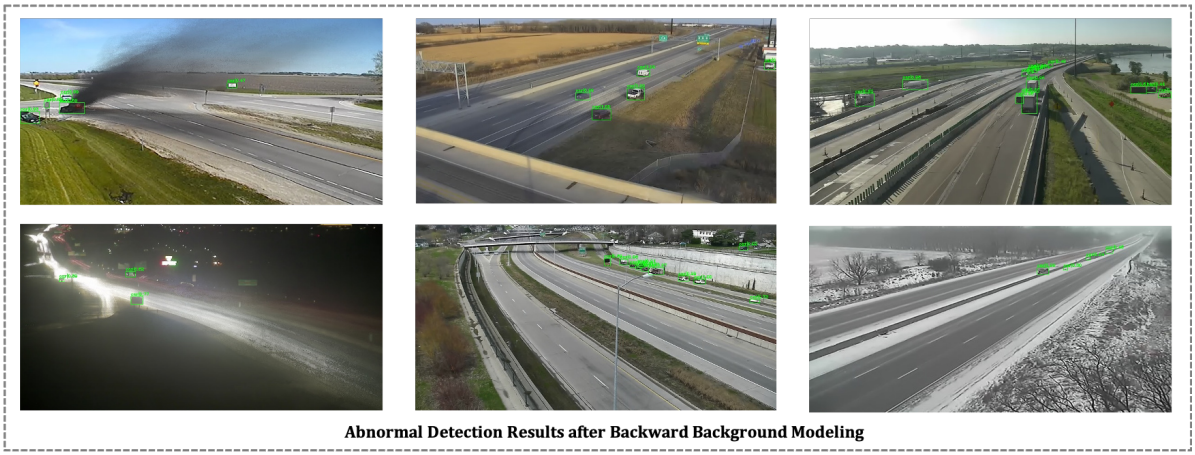

Figure 4. Examples of detection result after backward background modeling. From the figure we can observe that abnormal stopped car can be detected successfully.

hicles, which involves a box-level tube tracking process and a refinement scheme.

2.4.1 Tube Tracking

Box Linking and Tube Construction. We first perform the box linking and tube construction process. We follow [8] to generate box-level tracking results. Specifically, we first adopt the muiti-stage detection approach in section 2.1 to detect all bounding boxes, $\{B\}$ in the video frames after the backward background modeling process, with corresponding confidence scores $S(B)$. We first filter the detection results. If the intersection-over-union (IoU) between the detection box and the mask area is less than λ or the predicted confidence $S(B)$ is less than λ, we will discard these boxes. Subsequently, we link the detections across the single frame to produce a spatio-temporal tube to outline a particular vehicle. Then, we sort the detections according to $S(B)$ and pick the one with the max score as the starting point of the linking process. Then the linking process is extended both forward and backward via the greedy search algorithm and the box with the max linking score in the consecutive time is added to the corresponding tube. Specifically, the linking score $S_l(B_i, B_j)$ is defined as the intersection-over-union (IoU) of two detection boxes B_i and B_j. We continue the linking process until there is no box could obtain the IoU greater than λ. When a special tube is constructed, we remove the linked boxes and collect a new tube repeatedly until all boxes are grouped.

2.4.2 Tube Refinement.

In complex traffic scenes, the detection results are essential to the final results. To deal with the issue of false detection, we designed four mechanisms to compensate for the detection performance, i.e., spatial fusion, still-thing filter, temporal fusion and feedforward optimization.

Spatial Fusion. First, we compare the starting boxes of the candidate tubes. When their IoU exceeds the threshold λ, we consider that these two tubes are related to the same vehicle and merge these tubes into the same tube.

Still-thing Filter. [8] proposes a similarity filter to discard some false detections. However, they are complicated and have more hyper-parameters to adjust. In this paper, we introduce a still-thing filter module to exclude some background information. We employ detection results from the original video frame to guide this process. Specifically, we use the detection result of the original frame to confirm whether the candidate tube is false detection. In other words, if the original frame always has the similar detection result (IoU surpasses λ) outside the tube, it means that the object may be an error in the video that has always existed. Hence we consider they are actually background information and filter them out. We illustrate the motivation of the still-thing filter in Figure 5. (a) and (b) represents the backward background modeling detection results before and after the still-thing filter, figure (c) depicts the origin video detection results. As some still thing such as road sign (red elliptic region) may be falsely detected, still-thing appears along whole origin video will be treated as normal instance and will be filtered via this module.

Temporal Fusion. Then we merge the obtained tubes in the temporal dimension. When the end time of the current abnormal tube $t_e^{T_i}$ and the start time $t_s^{T_{i+1}}$ of the next one are within β, we think they belong to the same abnormal event and combine these tubes.

FeedForward Optimization. Considering that the results of background modeling in the backward direction may make the appearance of vehicles earlier, we additionally employ the detections of the original frames to refine

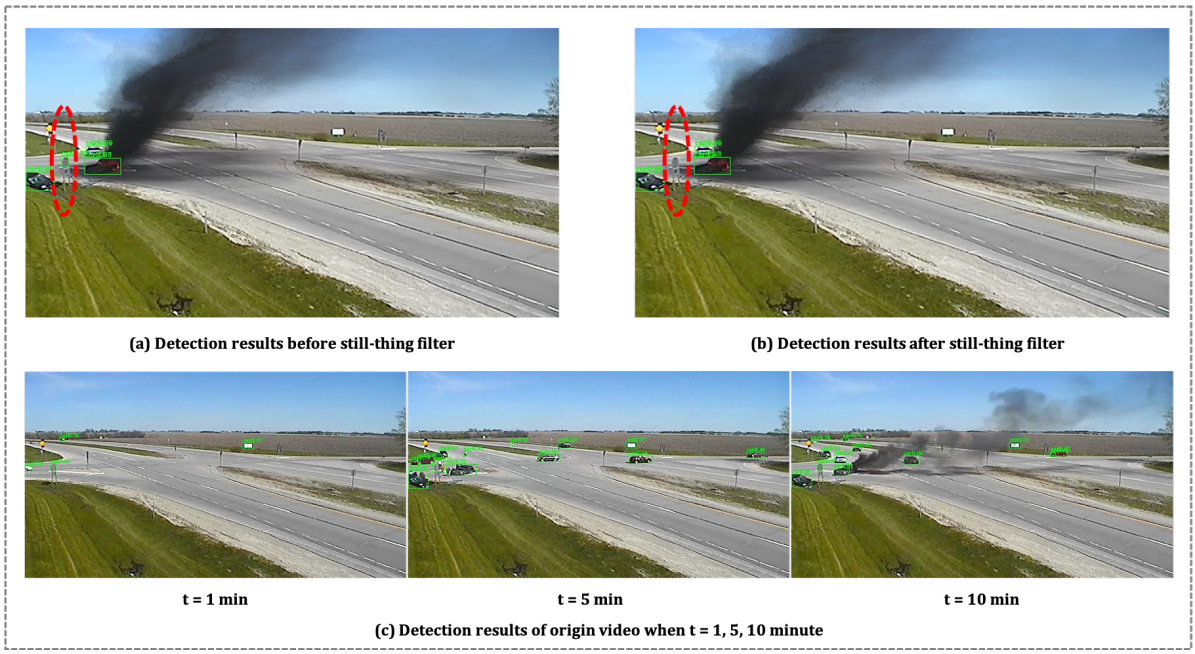

Figure 5. Example of still-thing filter. Figure (a) and (b) stand for backward background modeling detection results before and after the still-thing filter, figure (c) depicts the origin video detection results. As some still thing such as road sign (red elliptic region) may be falsely detected, still-thing appears along whole origin video will be treated as normal instance and will be filtered.

the abnormal results. Specifically, we use the detections of the start time of the candidate tube to compare with the detections of the original frames in the corresponding time. When the number of feedforward frames is less than the max feedforward frame ζ and the IoU between the detections is greater than the IoU threshold λ, we update the starting time of this anomaly to the time of the current detection. The feedforward process is repeated until the threshold condition is not met.

3. Experiments

3.1. Experimental Setup

The track4 dataset in NVIDIA AI CITY CHALLENGE 2021 is divided into the training set and test set. The training set contains 100 videos with a length of approximately 15 minutes, a frame rate of 30 fps and a resolution of 800 $\times$ 410. And the testing set consists of 150 videos. The algorithm should identify all car crashes or stalled vehicles in all 150 test set videos, and give the start time and the corresponding confidence score. We first conduct the experiments in the training set to determine the model parameters through cross-validation. Then we directly adopt the parameters of each component obtained by cross-validation to obtain the final result in the test set.

3.2. Implementation Details

Extraction of Differential Abnormal Mask. For the differential mask, the hyper parameter k is set to 5. Namely, we extract five frames per second to calculate the changing area. The difference threshold η is 99.

Backward Background Modeling. In the backward background modeling, we fix $T = 4$ to update the GMM parameters. In order words, the update interval is set as 120 frames at 30 fps. As a result, all normal moving vehicles are removed from the frames and static vehicles remain in the background.

Detection Model. We leverage Cascade R-CNN with the backbone of ResNeXt-101 (64x4d) pretrained on ImageNet as our detection model. FPN with 5 layer are used to build high-level semantic feature maps at all scales, the downsample strides are [4, 8, 16, 32, 64]. We use 3 stage cascade heads with different IoU thresholds 0.5, 0.6 and 0.7. We adopt Stochastic Gradient Descent (SGD) optimizer with momentum rate 0.9 and weight decay 1e-4 to train the model for 20 epochs. We employ 8 Nvidia Tesla V100 to accomplish the training process and set batch size to 32. The initial learning rate is 0.02 and it will be reduced by a factor of 10 at epoch 16 and 19. We follow the setting of popular detection framework[17, 2], setting the input size as 1333×800. We adopt random horizontal flip as the data augmentation in the train process and the random probability is set to 0.5. We fix the score threshold to 0.5 and the NMS IoU threshold to 0.2 during the inference process.

Table 1. Experimental results on Track4 test-set.

Team ID	F1	RMSE	Total Score
76	-	-	0.9355
158 (Our)	0.9318	3.1623	0.9220
92	-	-	0.9197
90	-	-	0.8597
153	-	-	0.5686

Box-level Tracking and Refinement. The linking IoU threshold λ is 0.4. The max traceback frame ζ is 240 frames. The temporal fusion threshold β is set to 5000 frames.

3.3. Evaluation Metric

A robust metric is adopted to measure the performance of anomaly detection, which is computed via F1-score and normalized root mean square error (NRMSE):

$$S4 = F1 \times (1 - \mathrm{NRMSE}). \tag{1}$$

The F1-score is the harmonic mean of recall and precision. Specifically, a true-positive (TP) detection is considered as the correct anomaly within ten seconds of a real anomaly. A false-negative (FN) is a real abnormal event that the proposed approach fails to correctly predict. A false-postive (FP) respents the predicted anomaly is not a real anomoly actually. The F1-score is summarized by:

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}. \tag{2}$$

Normalized root mean square error (NRMSE) denotes the temporal error of the predicted time and real anomaly time for all true-positive predictions. NRMSE employs a max-min normalization with a maximum value of 300 and a minimum value of 0. In short, NRMSE is defined as follow:

$$\mathrm{NRMSE} = \frac{\min(\sqrt{\frac{1}{\mathrm{TP}}\sum_{i=1}^{\mathrm{TP}}(t_i^p - t_i^{gt})^2}, 300)}{300}, \tag{3}$$

where t_i^{gt} denotes the ground truth starting time of the anomaly and t_i^p is the predicted starting time proposed by our approach.

3.4. Experimental results

We evaluate our method on the NVIDIA AI CITY CHALLENGE 2021 Anomaly Detection Track testing data. As shown in Table 1, we achieve 0.9318 F1-score while the start time error is only 3.1623 seconds, which demonstrates the superiority and robustness of our proposed method. We achieve 0.9220 S4 score and rank the second place among all the participant teams.

4. Conclusions

In this paper, we design a box-level tracking and refinement approach, which contains the extraction of hypothetical differential mask, backward background modeling to eliminate dynamic traffic disturbance, the multi-stage detection model to obtain still vehicles, a box-level tracking mechanism to construct candidate abnormal tubes , and a refinement scheme to promote a more accurate abnormal period. Results on NVIDIA AI CITY CHALLENGE 2021 show our proposed method shows promising performance, which gets a 0.9220 total score, 93.18% F1-score and 3.1623 RMSE.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3449–3456. IEEE, 2011.

[4] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016.

[5] Timothy Hospedales, Shaogang Gong, and Tao Xiang. A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1165–1172. IEEE, 2009.

[6] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009.

[7] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1453. IEEE, 2009.

[8] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 586–587, 2020.

[9] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE Inter-*

national Conference on Computer Vision, pages 2720–2727, 2013.

[10] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[11] Hajananth Nallaivarothayan, Clinton Fookes, Simon Denman, and Sridha Sridharan. An mrf based abnormal event detection approach using motion and appearance features. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 343–348. IEEE, 2014.

[12] Milind Naphade, Ming-Ching Chang, Anuj Sharma, David C Anastasiu, Vamsi Jagarlamudi, Pranamesh Chakraborty, Tingting Huang, Shuo Wang, Ming-Yu Liu, Rama Chellappa, et al. The 2018 nvidia ai city challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 53–60, 2018.

[13] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, et al. The 2019 ai city challenge. In *CVPR Workshops*, 2019.

[14] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 2665–2674, June 2020.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[16] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[17] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[18] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[19] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[20] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *IEEE International Conference on Image Processing*, pages 4030–4034. IEEE, 2019.

[21] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3320. IEEE, 2011.

[22] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019.

[23] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.

[24] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004.

[25] Zoran Zivkovic and Ferdinand Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):p.773–780, 2006.