Towards Real-Time Systems for Vehicle Re-Identification, Multi-Camera Tracking, and Anomaly Detection

Neehar Peri*[1] , Pirazh Khorramshahi*[1], Sai Saketh Rambhatla*[1], Vineet Shenoy[1], Saumya Rawat[1], Jun-Cheng Chen[2] and Rama Chellappa[1]
[1]Center for Automation Research, UMIACS, University of Maryland, College Park
[2]Research Center for Information Technology Innovation, Academia Sinica
{peri, pirazhkh, rssaketh, vshenoy, saumya, rama}@umiacs.umd.edu,
pullpull@citi.sinica.edu.tw

Abstract

Vehicle re-identification, multi-camera vehicle tracking, and anomaly detection are essential for city-scale intelligent transportation systems. Both vehicle re-id and multi-camera tracking are challenging due to variations in aspect-ratio, occlusion, and orientation. Robust re-id and tracking systems must consider small scale variations in a vehicle's appearance to accurately distinguish among vehicles of the same make, model, and color. Scalability is critical for multi-camera systems, as the number of objects in a scene is not known a-priori. Anomaly detection presents a unique challenge due to a dearth of annotations and varied video quality. In this paper, we address the task of vehicle re-id by introducing an unsupervised excitation layer to enhance representation learning. We propose a multi-camera tracking pipeline leveraging this re-id feature extractor to compute a distance matrix and perform clustering to obtain multi-camera vehicle trajectories. Lastly, we leverage background modeling techniques to localize anomalies such as stalled vehicles and collisions. We show the effectiveness of our proposed method on the NVIDIA AI City Challenge, where we obtain 7^{th} place out of 41 teams for the task of vehicle re-id, with an mAP score of 66.68% and achieve state-of-the-art results on the Vehicle-ID dataset. We also obtain an IDF1 score of 12.45% on multi-camera vehicle tracking, and an S4 score of 29.52% for task of anomaly detection, ranking in the top 5 for both tracks.

1. Introduction

In recent years, there has been great demand to develop automated and intelligent transportation systems for smart cities that can facilitate dynamic traffic routing, traffic planning, gathering vehicle-specific analytics like speed [17], and traffic anomaly detection. Moreover, the development of Deep Convolutional Neural Networks (DCNNs) has enabled the development of effective solutions to these challenges. For the past three years, NVIDIA AI City Challenge has pushed the boundaries of intelligent transportation systems. In this paper, we present a deep learning-based algorithm for the task of vehicle re-identification (re-id), and end-to-end pipelines for Multi-Camera Tracking (MTC) and anomaly detection.

Vehicle re-id refers to the task of identifying all true matches of a given vehicle identity in a large gallery set composed of images of different vehicles that are captured under diverse conditions, *e.g.*, different image quality, orientation, weather condition and lightening. Therefore, learning robust representations able to handle the aforementioned conditions is of great importance. At the same time, a representation learning algorithm should be both real-time and scalable to adapt to a large number of vehicles and traffic cameras in the wild. To this end, we propose the fast and accurate Excited Vehicle Re-identification (EVER) model to meet these challenges. Recent work has shown the importance of attending to local regions, vehicle key-points, [13, 37] and part bounding boxes [9] to create robust deep features. However, generating key-point annotations and part bounding boxes is costly and will not scale across different domains. [14] has proposed a novel self-supervised model to generate residual maps that act as pseudo-attention maps. In this work, we take advantage of the residuals generated from [14] to excite intermediate feature maps during the course of training and encourage the feature extraction model to learn robust representations.

Multi-Camera Tracking aims to determine the position of objects under consideration, at all times from video streams taken by multiple cameras. The resulting multi-camera trajectories enable applications including visual an-

*The first three authors equally contributed to this work.

alytics, suspicious activity and anomaly detection. In recent years, the number of cameras in highways, parking lots and intersections have increased dramatically, so it has become paramount to automate MCT. MCT is a notoriously difficult problem: Cameras are often placed far apart to reduce costs, and their fields of view do not always overlap. This results in extended periods of occlusion and large changes in viewpoint and illumination across different fields of view. In addition, the amount of data to process is enormous. In this work, we present a system for MCT that leverages advances in Single Camera Tracking [32,36,39] and our proposed vehicle re-id model discussed above to obtain trajectories of vehicles under different cameras.

Vehicle anomaly detection attempts to automatically localize stalled vehicles and collisions using existing traffic camera infrastructure. Anomalous vehicles are uniquely represented in both the foreground and background of a scene. Parked vehicles are typically only represented in a background model, while moving vehicles are only represented in the foreground model. Anomalous vehicles can be characterized by their transitions between the foreground and background. These transitions provide distinct opportunities to localize the spatio-temporal bounds of anomalous vehicles. Our proposed method leverages this property to identify a variety of anomalies, while also minimizing computational complexity. The proposed anomaly detection algorithm first creates background and foreground masks for each frame. We detect vehicles that are present in the background image, and filter these proposals using a pre-calculated road mask. It is important to note that this method is unsupervised, and uses a Hybrid-Task Cascade Network [3] pretrained on COCO [19] from the MMdetection framework [4]. We achieve an F1 score of 59.46%.

2. Related Works

Vehicle Re-identification: Successful Vehicle re-id requires learning features robust to variations in orientation, illumination and occlusion. Due to the expansive literature, we briefly review several recent methods on vehicle re-identification.

Large-scale vehicle re-id datasets such as Vehicle-ID [21], VeRi-776 [22], CityFlow-ReID [33] have made it possible to learn global feature embeddings. However, these global representations may fail to take into account the minute details among visually similar vehicles of the same make, model and color. In addition, the global appearance of a given vehicle varies significantly as its viewpoint changes depending on the camera. To alleviate this issue, several methods [9, 13, 23, 37, 41] have been proposed to enhance the discrimative capability of DCNNs by enforcing attention on local regions of a vehicle such as head and tail lights, grill, bumpers and wheel patterns. Zhou *et al.* [44] learns a viewpoint-aware representation for vehicle re-id through view-dependent attention. [14] proposed a self-supervised attention generation eliminates the need for extra annotations for vehicle's local regions. Also, [27,42] leverages vehicle attribute classification to attend to informative regions, *e.g.*, predicting color and vehicle type to learn attribute-based auxiliary features to assist the global representation. Metric learning is widely used in an effort to make robust representations. [5, 16] propose various triplet losses to carefully select hard triplets across different viewpoints and vehicles to learn an improved appearance-robust representation.

Multi-Object Tracking (MOT): Object tracking plays an important role in solving many fundamental computer vision tasks. The success of object detectors [6, 10, 18, 20, 28] has garnered significant interest in object tracking, resulting in many of robust single camera trackers for pedestrian and vehicles. Sun *et al*. [31] posed MOT as a data association problem and trained a Deep Affinity Network (DAN) to obtain the association matrix in an end-to-end fashion. DAN also accounts for multiple objects appearing and disappearing between video frames. [35] extends the problem of MOT to multi-object tracking and segmentation (MOTS). Voigtlaender *et al.* in [35] annotate dense pixel-wise labels for existing tracking datasets using a semi-automatic annotation procedure and propose a new baseline which jointly addresses detection, tracking and segmentation. [38] argues that the two-stage (object detection followed by data association) tracking-by-detection paradigm suggested by most modern MOT systems can lead to efficiency issues for real-time MOT and hence proposed a real-time system that facilitates learning detection and appearance embeddings by a shared model. They formulate the problem as a multi-task learning setup and report the first near real-time MOT system.

Multi-Camera Multi-Object Tracking: [30] learns good features for multi-camera tracking and re-id with a DCNN using an adaptive weighted triplet loss for training and a new technique for hard-identity mining. [34] proposed a unified three-layer hierarchical approach for solving tracking problems in multiple non-overlapping cameras. Similar to our proposed pipeline, Tesfaye *et al.* in [34] first solve within-camera tracking and then solve across-camera tracking by merging tracks of the same object in all cameras in a simultaneous fashion. They use the constrained dominant sets clustering (CDSC) technique, a parametrized version of standard quadratic optimization to solve both the tracking tasks.

Anomaly Detection: Recent work in vehicle anomaly detection has focused on detecting stopped vehicles through a two-stage pipeline, first using background modeling to identify vehicle proposals, and refining these proposals by identifying regions of interest. [1] models each scene by computing an moving-average image for each scene, de-

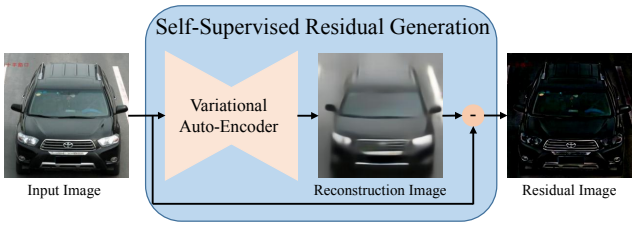

Figure 1. SAVER [14] crudely reconstructs the input image using a VAE. Subtracting the generated image from the input yields the residual image in which salient regions are highlighted.

tects multi-scale vehicles through a perspective transformation, and regresses the start and end time of each anomaly through a spatio-temporal information matrix. Similarly, [26] approaches the task of background modeling by calculating an average image for each scene, and utilizes multiple detectors optimized for various road conditions to localize anomalous vehicles. [15] considers the spatio-temporal consistency of tracklets to filter out moving vehicles and refines these predictions by constructing binary masks to highlight regions of interest. Our proposed method differs from previous approaches, and leverages a Gaussian mixture model (GMM) to simultaneously create background and foreground representations and identify anomalous vehicles in near real-time.

3. Vehicle Re-Identification

In this section we present our proposed approach, Excited Vehicle Re-identification (EVER), for the vehicle re-id track of the challenge. EVER consists of three modules, namely Self-Supervised Residual Generation, an Excitation Layer and Feature Extraction.

3.1. Self-Supervised Residual Generation

Inspired by SAVER [14], which generates both a per image coarse template of a given vehicle and a residual image that carries vehicle-specific details critical for re-id, we take advantage of the residual image to generate robust features. The residual image containing minute details serves as a pseudo-attention map. Figure 1 demonstrates how a residual image is obtained and how it highlights salient parts of the vehicle.

3.2. Excitation Layer

Although SAVER uses the residuals to augment the input image to a re-id model via convex combination, we propose to only employ the residuals during training to excite the intermediate feature maps and assist the feature extraction model to learn more discriminative vehicle representations. Intermediate feature map excitation has been shown to be an effective approach for vehicle re-identification [9, 13]

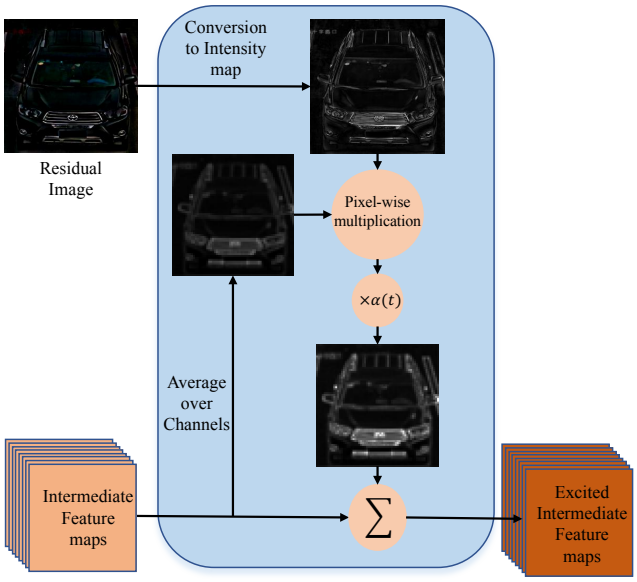

Figure 2. Intermediate feature maps are excited during training with the help of the residual image. This is done by producing an excitation that is a function of the residual image, channel-averaged feature maps and $\alpha(t)$. Note that during inference this layer acts as an identity function on the feature maps.

and object detection [7]. In particular, we follow the excitation method proposed in [7] which is only applied significantly during initial epochs of training and monotonically decreases the degree of excitation as training continues. This is done by computing the excitation factor $\alpha(t)$ as follows:

$$\alpha(t) = 0.5 \times \left(1 + \cos(\frac{\pi t}{T})\right) \quad (1)$$

where $t = 1 \ldots T$ is the epoch number and T is the total number of training epochs. Figure 2 shows how intermediate feature maps are excited while training the re-id model. Subsequently, inference is only requires a forward pass of the re-id model without generating residual image, *i.e.*, $\alpha(t) = 0$, which ultimately reduces the computational complexity of EVER. This makes our proposed approach quite competitive for real-time applications.

3.3. Feature Extraction

For the purpose of discriminative deep feature extraction, we chose the backbone architecture of ResNet-152 [11] for our re-id model. Recently, [24] established a set of training techniques for the task of person re-id which is shown to outperform many complicated methods and serves as a strong baseline. These tricks have also been shown as effective for the task of vehicle re-id [14]. Therefore, as our baseline, we adopt these tricks, *i.e.*, Learning Rate Warm-up, Random Erasing Augmentation, Label Smoothing, and Batch Normalization Neck, for training the ResNet model

Table 1. Performance comparison between baseline and the proposed method on Large-scale Vehicle Datasets

Model	Dataset											
	CityFlow-ReID			Veri-776			VehicleID					
	mAP(%)	CMC(%)		mAP(%)	CMC(%)		Small		Medium		Large	
							CMC(%)		CMC(%)		CMC(%)	
		@1	@5		@1	@5	@1	@5	@1	@5	@1	@5
Baseline	62.36	60.55	60.74	78.51	95.10	98.00	80.00	95.30	77.30	91.70	75.20	88.60
Proposed	**66.68**	**65.40**	**65.68**	**79.90**	**95.90**	**98.20**	**84.50**	**96.40**	**79.70**	**94.70**	**77.40**	**91.80**

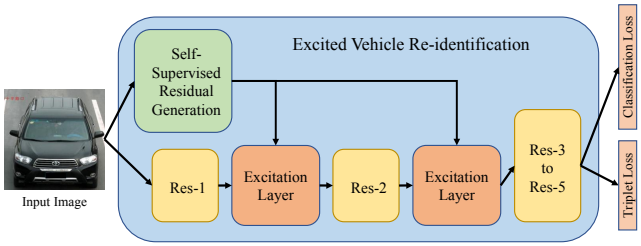

Figure 3. Overview of proposed EVER pipeline. Re-id model has ResNet-152 architecture. During training features maps after Res-1 and Res-2 blocks are excited with the help of residuals.

and compare it against similar settings with the addition of the excitation layer, *i.e.* EVER. Figure 3 shows our proposed pipeline. We optimized both baseline and EVER models for the following batch hard triplet [12] and cross entropy objectives:

$$\mathcal{L}_t = \frac{1}{B}\sum_{i=1}^{B}\sum_{a\in b_i}\left[\gamma + \max_{p\in\mathcal{P}(a)}||x_a - x_p||_2 - \min_{n\in\mathcal{N}(a)}||x_a - x_n||_2\right]_+ \tag{2}$$

and

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\left[(W_{c(x_i)}^T x_i + b_i) - \log\left(\sum_{j=1}^{C} e^{W_j^T x_i + b_j}\right)\right] \tag{3}$$

In Eq. 2, B, b_i, a, γ, $\mathcal{P}(a)$ and $\mathcal{N}(a)$ are the total number of batches, i^{th} batch, anchor sample, distance margin threshold, positive and negative sample sets corresponding to a given anchor respectively. Moreover, x_a, x_p, x_n are the extracted features for anchor, positive and negative samples. Batches are constructed in a way that they have exactly 4 instances of each ID used. In Eq. 3, x_i refers to the extracted feature for an image belonging to class i. Furthermore, $W_{c(x_i)}$, b_i are the classifier's weight vector and bias associated with class i respectively, and N and C represent the total number of samples and classes in the training dataset.

3.4. Experiments

In this section, we test the effectiveness of our proposed method. We compute the most commonly used re-id metrics, namely mean Average Precision (mAP) and Cumulative Matched Cure (CMC) @1 and @5 for CityFlow-ReID, Veri-776 and VehicleID benchmarks. Table 1 compares the performance of both the baseline and EVER models. The evaluation of CityFlow-ReID is done via an online server intended for the Challenge. With the goal of achieving the highest performance among participating teams, we applied re-ranking method [43] on our model's extracted features from CityFlow-ReID dataset as a post-processing step. It can be observed that for all the three datasets EVER model significantly improves the re-id metrics and achieves the state-of-the-art results on all three test splits of VehicleID dataset. Table 2 shows how our model is ranked among top performers of the challenge.

Table 2. Top 8 performers of 2020 NVIDIA AI City vehicle re-id challenge

Team Name	**mAP (%)**
Baidu-UTS	84.13
RuiYanAI	78.10
DMT	73.22
IOSB-VeRi	68.99
BestImage	66.84
BeBetter	66.83
UMD_RC	**66.68**
Ainnovation	65.61

3.5. Run-time performance

As discussed in section 3.2, one of the main advantages of EVER is its inference run-time. On a single GeForce TITAN Xp card, on average it takes only 13.5 milliseconds to process batches of size 128 and extract robust features. This makes EVER a particularly fast model for real-time applications.

4. Multi-Camera Tracking

In this section we describe our Multi-Camera Tracking (MCT) pipeline. We start by describing various Single-Camera Tracking (SCT) methods used in our work in Section 4.1. We then describe our Multi-Camera Tracking pipeline in Section 4.2. We conclude by describing all experiments in Section 4.3.

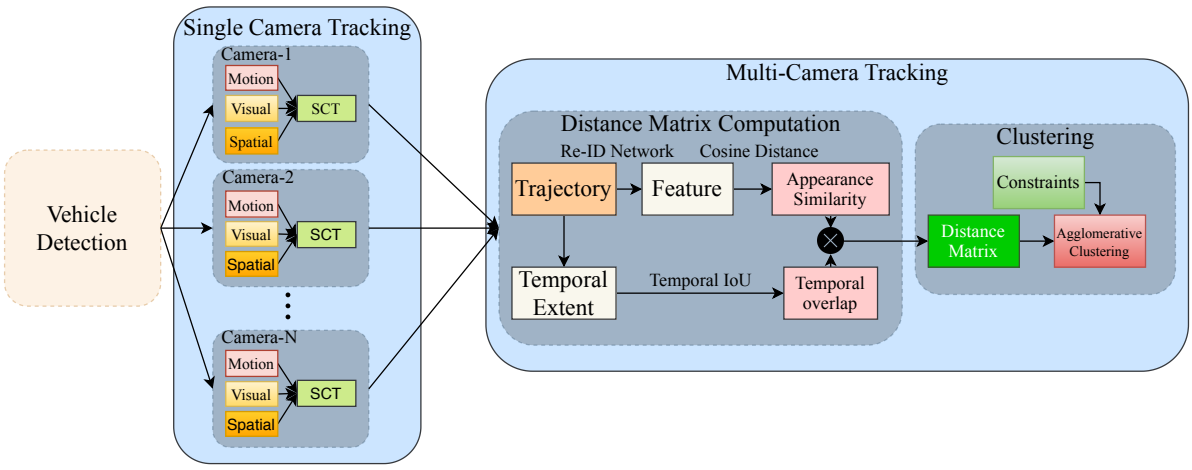

Figure 4. Proposed pipeline for Multi Camera Tracking of Vehicles

4.1. Single-Camera Tracking

Significant advances in object detection [6, 10, 18, 20, 28] aided the emergence of the tracking-by-detection paradigm [8, 40], which drastically improved the performance of various SCT methods for human and vehicle tracking. Such methods leverage the highly accurate spatial localization capabilities of the detectors, along with well-embedded appearance and temporal relationships for computing similarity measures to determine accurate object tracks.

4.1.1 Object Detection

In this work, we use Mask-RCNN object detector proposed in [10]. Mask R-CNN network builds on the Faster R-CNN [28] architecture with two major contributions. 1) Replacing the ROI Pooling module with a more accurate ROI Align module and 2) Inserting an additional branch (other than classification and bounding box heads) out of the ROI Align module to compute the object mask for the task of instance segmentation.

4.1.2 DeepSORT

Simple online and real-time tracking (SORT) [2] is a simple framework that performs Kalman filtering in image space and frame-by-frame data association using the Hungarian method with an association metric that measures the bounding box overlap. While achieving overall good performance in tracking precision and accuracy, SORT returns a relatively high number of identity switches. To improve the performance of SORT, DeepSORT [39] was proposed to overcome the issue by replacing the original association metric with a metric that combines motion and appearance information from a CNN pre-trained on a re-id dataset.

4.1.3 MOANA

Due to noisy detections and identity switches caused by occlusion and similar appearance among nearby targets in MOT systems, [32] proposed "Modeling of Object Appearance by Normalized Adaptation"(MOANA) that learns online a relatively long-term appearance change of each target. The proposed model is compatible with any feature of fixed dimension or its combination, whose learning rates are dynamically controlled by adaptive update and spatial weighting schemes. To handle occlusion and nearby objects sharing similar appearances, they design cross-matching and re-id schemes based on the application of the proposed adaptive appearance models.

4.1.4 TrackletNet Tracker (TNT)

TNT [36] leverages appearance, temporal and interaction cues together into a unified framework based on an undirected graph model. The vertices in the graph model are tracklets and the edges measure connectivity of two tracklets. Under such a graphical representation, tracking can be regarded as a clustering problem that groups the tracklets into one big cluster. The tracklets are generated based on IoU and appearance features similarity. When these criteria become unreliable due to camera motion, they adopt epipolar geometry to compensate and predict the position of bounding boxes in the next frame. TNT is trained to measure the continuity of two input tracklets by combining both trajectory and appearance information.

4.2. Multi-Camera Tracking Pipeline

Our Multi-Camera tracking pipeline proceeds as follows

- Detect and track all vehicles in all the videos.
- Extract EVER re-id features from every track to use as track descriptors

Table 3. Comparison of 3 SCT algorithms on 4 videos of 2 scenes in the validation set provided by the NVIDIA AI City Challenge 2020.

Tracking Method	S02				S05			
	c006	c007	c008	c009	c010	c016	c017	c018
DeepSORT	6.4	43.3	11.2	16.1	10.9	39.4	21.9	63.3
MOANA	8.7	**51.8**	**15.7**	**21.4**	**11.2**	38.6	**34.6**	67.8
TrackletNet	**9.3**	50.0	15.3	17.8	11.1	**43.4**	25.6	**71.5**

Table 4. Comparison of MCT algorithms on 4 videos of S02 in the validation set of the 2020 NVIDIA AI City Challenge.

Tracking Method	S02		
	IDF1	IDP	IDR
DeepSORT	44.13	63.57	33.80
MOANA	28.43	33.87	24.50
TrackletNet	33.19	41.12	27.83

Table 5. Top 8 performers of 2020 NVIDIA AI City multi-camera tracking challenge

Team Name	**Score**
INF	0.4585
XJTU-Alpha	0.4400
DukBaeGi	0.3483
EINI_CQUPT	0.3411
UMD_RC	**0.1245**
Albany_NCCU	0.0620
Youtu	0.0452
SJTU_yutinggao	0.0387

- Construct a distance matrix using appearance and temporal cues
- Cluster all the tracks to obtain final multi-camera tracks

The overall system is shown in Figure 4.

Single Camera Tracks: We use a SCT to get complete vehicle tracks for every video.

Track Descriptors: Owing to the superior discriminative ability of our proposed EVER (Section 3) system, we use it to extract re-id features for N ($N = 10$ in this work) randomly selected frames for every track in all the videos to obtain the corresponding track descriptors. We use the track descriptors to compute a distance matrix $\mathcal{D}$ which can be used to merge tracks of vehicles under different cameras. Since models for vehicle re-id are trained to identify cars under different viewpoints and imaging conditions, it is fitting to use a re-id model to merge tracks from different cameras.

Distance matrix using Appearance and Temporal Cues: Using the track descriptors, we compute a distance matrix $\mathcal{D} = [d_{ij}]_{i,j=0}^{i,j=M} = 1 - cos(f_i, f_j)$, where $cos(u, v) = \frac{u^T v}{\|u\|\|v\|}$ is the cosine similarity between vectors u, v; f_i, f_j are track i and j descriptors respectively and M is the total number of tracks from all the videos. Furthermore, two adjacent tracks of the same car usually have similarities in time. To incorporate this into the distance matrix, we use a temporal IoU. Specifically, we scale the distance matrix by temporal overlap in the following manner: $d_{ij} = d_{ij} * (1 - \frac{t_1 \cap t_2}{t_1 \cup t_2})$.

Clustering: After computing the distance matrix as described above, we perform clustering to obtain multi-camera tracks. Since tracks from the same camera shouldn't be merged together, we set the corresponding values in the distance matrix to a very high value to discourage the clustering algorithm to place the tracks in the same cluster. Since the number of clusters is not known beforehand, we apply bottom up Agglomerative clustering method to merge and obtain the multi camera tracks.

4.3. Experiments

Dataset: For all our experiments, we use the data provided as a part of 2020 NVIDIA AI City Challenge. The dataset contains 215.03 minutes of videos collected from 46 cameras spanning 16 intersections in a mid-sized U.S. city. The dataset is divided into 6 scenarios. 3 of the scenarios are used for training, 2 are used for validation, and the remaining one is for testing. In total, the dataset contains nearly 300K bounding boxes for 880 distinct annotated vehicle identities. Only vehicles passing through at least 2 cameras have been annotated.

Evaluation Metric: For MTMC tracking, the IDF1 score [29] will be used to rank the performance of each tracker. IDF1 measures the ratio of correctly identified detections over the average number of ground-truth and computed detections. Other popular evaluation measures adopted by the MOT challenge [25], such as Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), mostly tracked targets (MT), and false alarm rate (FAR) are also provided by the evaluation server. In Table 3 shows the comparison of the three SCT methods on scene 2 of the validation set. We use the IDF1 score for comparison as this is the metric used for ranking various submissions in the competition. In Table 4, we compare the multi-camera tracking performance for the three SCT methods. In Table

5, we compare the results of top 8 submissions in the public leaderboard of the 2020 NVIDIA AI City Challenge. For this submission we use the TNT [36] SCT.

5. Anomaly Detection

In this section, we present our approach for near real-time anomaly detection using statistical methods and out-of-the-box detection and tracking algorithms. Our method leverage a Gaussian mixture model to model both background and foreground instances, and uses a Hybrid Task Cascade Network and SORT for object detection and tracking respectively.

5.1. Foreground and Background Model

We use a Gaussian mixture based segmentation algorithm proposed by [45] that adaptively selects an appropriate number of Gaussian distributions for each pixel, and has been shown to adapt well to scenes with varying illumination. For each frame, we generate both foreground and background images. The Gaussian mixture model considers the last N frames when defining the background and foreground regions. Through experimental evaluation, we found that $N = 120$ adequately filters moving traffic, while capturing anomalous vehicles transitioning from the foreground to the background.

5.2. Vehicle Detection

Scale invariance and robustness to low resolution vehicle images are important considerations when selecting a vehicle detector for anomaly detection. We found that the Hybrid Task Cascade Network [3] is able to reliably localize small vehicles at low detector thresholds. Since running the Hybrid Task Cascade Network is computationally expensive, we only run detections on every 30^{th} frame, allowing our pipeline to run in near real-time. Furthermore, we only run the detector on background frames to reduce the number of occlusions.

5.3. Tracking

We utilize SORT as defined in 4.1.2, Since we only calculate detections on background frames, the SORT tracker drops tracks and reassigns identities less frequently. We use the length of the track as a proxy for the likelihood that a given track is anomalous. We avoid using deep-learning based trackers since it adds additional computational complexity to our proposed pipeline, and will likely only provide marginal benefit since a re-id model trained on high quality vehicle images will likely fail to generalize to this domain.

5.4. Post-Processing

Aberrations such as aliasing and frozen frames can introduce artifacts into all sub-systems in an anomaly detection pipeline. To avoid false positive predictions due to poor video quality, we detect when consecutive frames have a per-pixel difference less than a fixed threshold, and ignore predictions from that region of a video. Additionally, we construct a road mask to highlight regions of interest and remove false positive detections by averaging the foreground frames together.

5.5. Anomaly Detection Pipeline

Localizing anomalous vehicles in near real-time requires robust background modeling, object detection, and lightweight tracking. Anomalous vehicles, particularly stalled vehicles, uniquely transition between the foreground and background. A lack of supervised data leads to the use of traditional computer vision and statistical methods. Figure 5 demonstrates our end-to-end pipeline. We approach background modeling through the use of a Gaussian mixture model. Since vehicle detection is an essential part in localizing anomalies, we prioritized using a computationally expensive deep learning model, but only run inference on every 30^{th} frame. We apply an online tracker to cluster these detections, and apply several heuristics to remove false positive results.
Figure 6 shows the qualitative performance of our system on the 2020 NVIDIA AI City Challenge dataset. We note that our proposed pipeline is able to accurately spatially localize most anomalous vehicles. However, our pre-trained detector often produces false positive results in night scenes, and bad weather conditions. Additionally, our foreground model is able to produce high quality road masks in busy scenes, but creates sparse representations when vehicles in a scene are sparse, or moving very slowly.

5.6. Experiments

Dataset: The NVIDIA AI City Challenge provides 200 fixed camera videos of unconstrained traffic scenes taken from highways and intersections in Iowa. These 200 videos are divided equally into training and testing sets. Each video is approximately 15 minutes long. The dataset includes a variety of illumination and weather conditions.
Evaluation Metric: Each anomaly detection pipeline is evaluated on its ability to accurately localize the start time of an anomaly. The S4 score is defined as F1 $\times$ (1 $-$ NRMSE), where the F1 score is the accuracy in selecting videos containing anomalies, and NRMSE measures the accuracy of the temporal bounds for each prediction. Table 6 shows how our method compares against other top performers in the challenge.

5.7. Run-time Performance

We compute the run-time of each module in our proposed pipeline and show that end-to-end pipeline runs in near real-time. We run each module five times on five

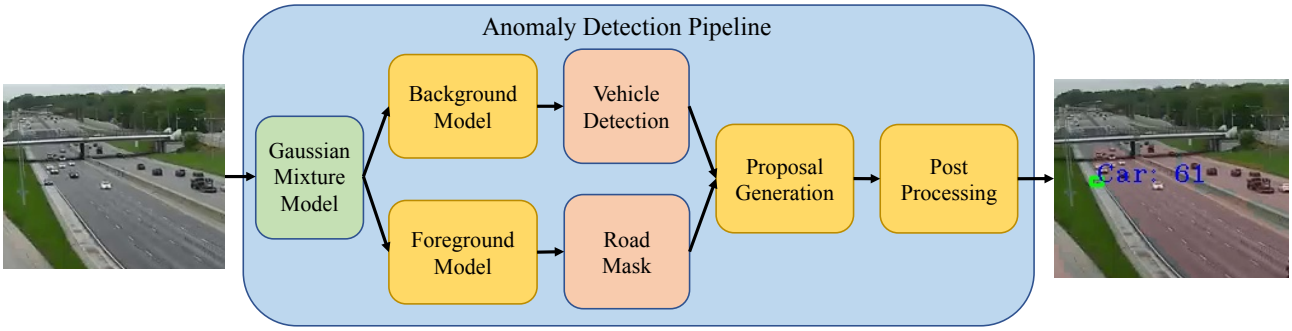

Figure 5. Our proposed anomaly detection pipeline is able to accurately localize multi-scale anomalies in near real time.

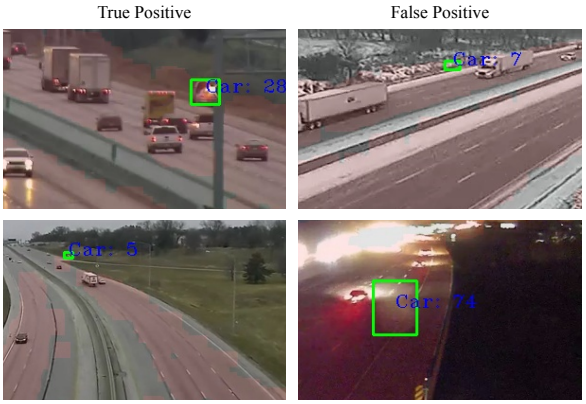

Figure 6. The Hybrid Task Cascade Network is able to accurately localize small vehicles in normal weather conditions, but generates false positive results in poor weather and less lighting.

Table 6. Top 8 performers in the 2020 NVIDIA AI City anomaly detection challenge

Team Name	S4 (%)
Firefly	96.95
SIS Lab	57.63
CETCVLAB	54.38
UMD_RC	**29.52**
HappyLoner	29.09
Orange-Control	23.86
PapaNet	0.1703
Team_Gaze_NSU_UAP	0.0958

videos each and average across each trial to normalize for variations in a given scene. All modules run on a CPU, except detector inference, which uses a single NVIDIA Titan X (Pascal). Table 7 demonstrates that the primary bottlenecks in our pipeline are the Gaussian mixture model and detector. Our current pipeline can process approximately 18 FPS. We can significantly reduce processing time by streaming relevant data to subsequent modules rather than saving to disk, and use a lighter detector trained on more domain specific data.

Table 7. Processing Time Analysis for 15-minute Video Clip

Component	Processing Time (minutes)
GMM Segmentation	12.7
Object Detection	11.39
Road Mask Construction	1.13
Object Tracking	0.05
Proposal Filtering	4e-6
Proposal Refinement	4e-6
End-to-End	25.29

6. Conclusion

In this paper, we summarizes our contributions to the 2020 NVIDIA AI City Challenge for the tasks of vehicle re-identification, multi-camera vehicle tracking, and anomaly detection, and highlight the computational efficiency of our proposed methods. As a byproduct, We achieve state-of-the-art results on the VehicleID dataset using the proposed EVER model and are ranked 7^{th} out of 41 teams. We are also ranked in the top 5 in public leaderboards for both multi-camera tracking and anomaly detection.

7. Acknowledgement

This research is supported in part by the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative, It is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 117–124. Computer Vision Foundation / IEEE, 2019.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016.

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4974–4983. Computer Vision Foundation / IEEE, 2019.

[4] 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.

[5] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8282–8291, 2019.

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 379–387. Curran Associates, Inc., 2016.

[7] Mohammad Mahdi Derakhshani, Saeed Masoudnia, Amir Hossein Shaker, Omid Mersa, Mohammad Amin Sadeghi, Mohammad Rastegari, and Babak N Araabi. Assisted excitation of activations: A learning technique to improve object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9201–9210, 2019.

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. *CoRR*, abs/1710.03958, 2017.

[9] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6132–6141, 2019.

[14] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. *arXiv preprint arXiv:2004.06271*, 2020.

[15] Pirazh Khorramshahi, Neehar Peri, Amit Kumar, Anshul Shah, and Rama Chellappa. Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 239–246. Computer Vision Foundation / IEEE, 2019.

[16] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019.

[17] Amit Kumar, Pirazh Khorramshahi, Wei-An Lin, Prithviraj Dhar, Jun-Cheng Chen, and Rama Chellappa. A semi-automatic 2d solution for vehicle speed estimation from monocular videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[18] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[20] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[21] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.

[22] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *IEEE International Conference on Multimedia and Expo, ICME*, pages 1–6. IEEE Computer Society, 2016.

[23] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. Ram: a region-aware deep model for vehicle re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[25] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[26] Khac-Tuan Nguyen, Trung-Hieu Hoang, Minh-Triet Tran, Trung-Nghia Le, Ngoc-Minh Bui, Trong-Le Do, Viet-Khoa Vo-Ho, Quoc-An Luong, Mai-Khiem Tran, Thanh-An Nguyen, Thanh-Dat Truong, Vinh-Tiep Nguyen, and Minh N. Do. Vehicle re-identification with learned representation and spatial verification and abnormality detection with multi-adaptive vehicle detectors for traffic video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 363–372. Computer Vision Foundation / IEEE, 2019.

[27] Jingjing Qian, Wei Jiang, Hao Luo, and Hongyan Yu. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. *arXiv preprint arXiv:1910.05549*, 2019.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[29] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[30] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6036–6046, 2018.

[31] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[32] Zheng Tang and Jenq-Neng Hwang. Moana: An online learned adaptive appearance model for robust multiple object tracking in 3d. *IEEE Access*, 7:31934–31945, 2019.

[33] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David C. Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8797–8806. Computer Vision Foundation / IEEE, 2019.

[34] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. *Int. J. Comput. Vis.*, 127(9):1303–1320, 2019.

[35] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe. Mots: Multi-object tracking and segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7934–7943, 2019.

[36] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet, 2018.

[37] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 379–387, 2017.

[38] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2019.

[39] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.

[40] H. Zhang, A. Geiger, and R. Urtasun. Understanding high-level semantics by modeling traffic patterns. In *2013 IEEE International Conference on Computer Vision*, pages 3056–3063, 2013.

[41] Xinyu Zhang, Rufeng Zhang, Jiewei Cao, Dong Gong, Mingyu You, and Chunhua Shen. Part-guided attention learning for vehicle re-identification. *arXiv preprint arXiv:1909.06023*, 2019.

[42] Aihua Zheng, Xianmin Lin, Chenglong Li, Ran He, and Jin Tang. Attributes guided feature learning for vehicle re-identification, 2019.

[43] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

[44] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 6489–6498, 2018.

[45] Zoran Zivkovic and Ferdinand van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognit. Lett.*, 27(7):773–780, 2006.