Attention Driven Vehicle Re-identification and Unsupervised Anomaly Detection for Traffic Understanding

Pirazh Khorramshahi, Neehar Peri, Amit Kumar, Anshul Shah and Rama Chellappa
Center for Automation Research , UMIACS
University of Maryland, College Park
{pirazhkh, peri, akumar14, rama}@umiacs.umd.edu, anshulb@cs.umd.edu

Abstract

Vehicle re-identification and anomaly detection are useful tools in traffic analytics applications. Vehicle re-identification is particularly challenging due to variations in viewpoint, illumination and occlusion. Moreover, the reality of multiple vehicles having the same make and model hinders the design of traditional deep network-based solutions. In this work, we leverage an attention-based model which learns to focus on different parts of a vehicle by conditioning the feature maps on visible key-points. We use triplet embedding to reduce the dimensionality of the features obtained from the ensemble of networks trained using different datasets. To address the problem of anomaly detection, we design an unsupervised algorithm to detect and localize anomalies in traffic scenes. To handle moving cameras, we use the results obtained from tracking to generate anomaly proposals which are then filtered in successive steps. We show the effectiveness of our method on the Nvidia AI City vehicle re-identification dataset, where we obtain mean Average Precision (mAP) score of 60.78% placing us at the 8th position out of 84 participating teams. In addition, we achieved the S3 score of 22.07% for vehicle anomaly detection.

1. Introduction

In the age of automation, there is a great need for automatic vehicle identification. In addition, it is also important to detect anomalous situations, such as stalled vehicles so that road side assistance can be dispatched immediately in an automated manner. Tasks such as re-identification of vehicles and detection of anomalous vehicles are important tools in traffic analysis for smart cities. In this work, we present a deep learning-based supervised method for vehicle re-identification and an unsupervised method for detecting anomalous vehicles.

Vehicle re-identification refers to the task of recognizing vehicles across different cameras placed in different locations and captured at different times. In an urban scene, this presents a great challenge as there can be a large number of vehicles having the same color, make and model. Furthermore, images are taken from different viewpoints which necessitates that systems be robust to variations in orientation. Our proposed method conditions the extracted features on the orientation of the vehicle through the localization

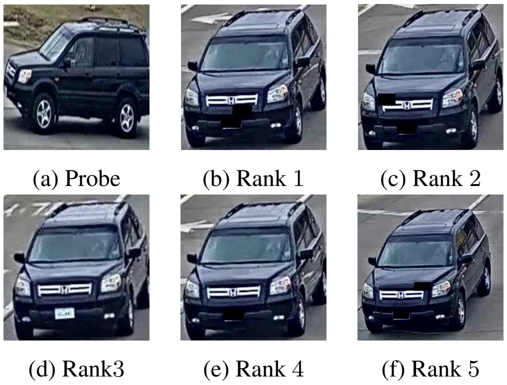

(a) Probe (b) Rank 1 (c) Rank 2
(d) Rank3 (e) Rank 4 (f) Rank 5

Figure 1: Successful retrieval by the proposed method for a probe image from CityFlow-ReID test data. All returned images appear to have the same identity as the probe.

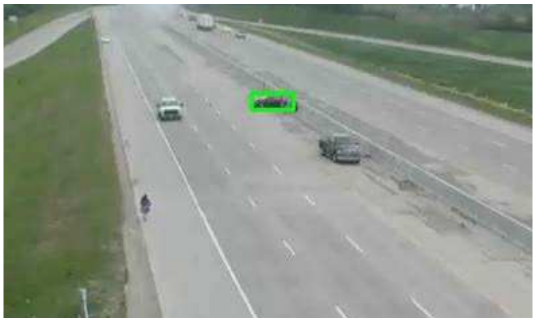


Figure 2: Successful detection of an anomaly using our proposed approach

of a vehicle's key-points. To reduce the dimensionality of the extracted features, we also use triplet probabilistic embedding (TPE) proposed by Sankaranarayanan *et al.* [19], which increases the speed of search and retrieval in a large gallery. Another aspect of the presented work is the use of an ensemble of networks which improves performance as different models trained with different datasets capture diverse discriminative features.

Anomalous vehicles are usually stalled for a longer period of time as compared to other vehicles in the scene. While many existing works make use of background subtraction as a method for isolating stalled vehicles, we use the results obtained from a vehicle tracker. This is done in order to handle the unstabilized videos acquired by moving cameras and dynamic scenes. The proposed algorithm first identifies vehicles which are on the street by creating a road mask. Next the vehicles on the periphery of the roads are analyzed by considering the vehicles around it in an adaptive window size. It is noteworthy to mention that we do not collect any labels and the proposed algorithm does not train any new neural networks. We only use a pre-trained Mask RCNN [4] for object detection, and the general appearance model from AAVER [6]. We call this method Context Aware Vehicle Anomaly Detection. With the scores obtained from the evaluation server we observe that the proposed method is able to achieve an F1 score of 57.14%.

2. Related Work

Since there exist significant intra and inter-class appearance variations (*i.e.*, pose, color, design, etc.) for different vehicles, learning an effective and discriminative feature representation is the key challenge for the Vehicle ReID task. We briefly discuss some recent relevant works in this section.

Different large scale datasets have been published in recent years including datasets for vehicle model classification and re-identification. Some of these datasets include CompCars by Yang *et al.* [29] for vehicle classification, VeRi [8, 9, 11] and Vehicle-ID by Liu *et al.* [10, 11] for vehicle re-identification. However, in comparison to Nvidia AICity dataset [24], the above datasets are relatively less challenging as the images contain little orientation variation and almost no occlusion.

In recent years, vehicle re-identification has gained momentum and significant progress has been made. Whereas Tang *et al.* [23] claimed the traditional hand-crafted features are complementary to deep features and proposed to fuse both features, Cui *et al.* [2] fused the features from various deep networks trained with different tasks and architectures. Concurrently, in [10, 11] Liu *et al.* proposed a coarse to fine approach by using multi-modal features, including visual features, license plate, camera location, and other contextual information. [30] and [28] used generative adversarial networks to synthesize vehicle images with diverse pose and appearance in order to augment the training data. In addition, both works demonstrated significant performance improvement for the vehicle ReID task. Besides the global features, Liu *et al.* [12] suggested region-aware deep model which extracted discriminative local features from a series of local regions of a vehicle. Similarly, Wang *et al.* [25] also proposed an orientation-invariant feature embedding (OIFE) to fuse the global feature and orientation-aware features which are generated by focusing on important local regions through orientation-aware region proposal. Recently, AAVER[6] proposed an adaptive attention obtained by considering the visible landmarks.

With increasing interest in public safety, anomaly detection has been studied in recent years. However, it has been limited to the context of detecting pedestrians on streets[16, 20]. Sabokrou *et al.* in [18] proposed a cascade classifier-based method to classify image patches into background or anomalies. [15] modeled human behaviour on streets using mathematical equations applied to pre-determined heuristics. In [22], to avoid labeling the segments of clips as anomalous, authors proposed using weakly labeled training samples in which videos with anomalies are labeled as anomalous. In contrast, our approach is unsupervised and models the behaviour of vehicles on streets by observing the tracklets and the movement of surrounding vehicles.

3. Track 2: City-Scale Multi-Camera Vehicle Re-Identification

The purpose of this track is to design a system that can find true matches to a given query image in a large gallery set. In the following subsections, we describe the designed dataset for this track. This is followed by a description of the proposed approach to tackle this problem.

3.1. Dataset

For this track a dataset, CityFlow-ReID [24], composed of 56,277 images of 666 different vehicle identities has been provided. The dataset is divided into 2 splits:

- **Training Split** : 36,935 images are considered for the this split. These images are gathered by tracking 333 vehicles in videos which generated 1897 tracks.
- **Testing Split**: 18,920 images of the remaining 333 vehicles captured in 798 tracks from the gallery set. The remaining 1052 images of the test split are used as the query or probe set.

3.2. Proposed Approach

The proposed vehicle re-identification approach is composed of three main stages: (1) Pre-Processing, (2) Discriminative Feature Extraction and (3) Post-Processing. Fig. 3

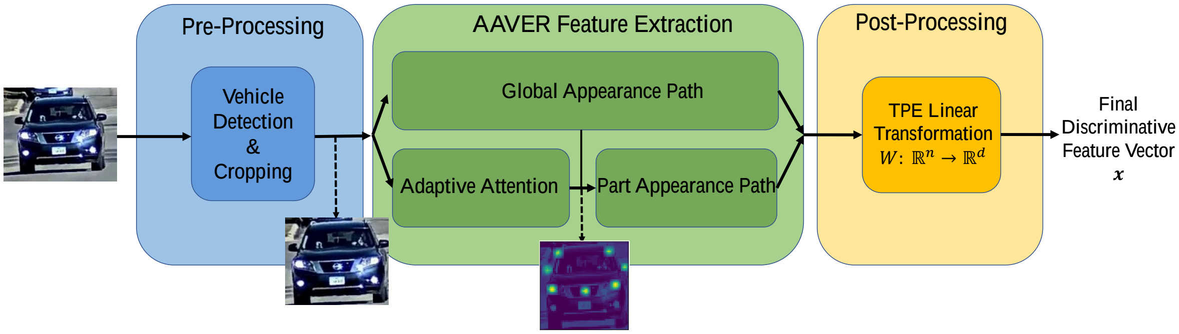
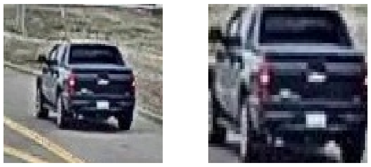

Pre-Processing
Vehicle Detection & Cropping
AAVER Feature Extraction
Global Appearance Path
Adaptive Attention
Part Appearance Path
Post-Processing
TPE Linear Transformation
$W: \mathbb{R}^n \rightarrow \mathbb{R}^d$
Final Discriminative Feature Vector
$\boldsymbol{x}$

Figure 3: The proposed approach for robust feature extraction pipeline. The input image is initially cropped to discard unwanted parts of the image. Features from the global appearance of the vehicle along with local features obtained through adaptive attention [6] are fused. Finally, the resulting feature vector is mapped to a lower dimensional space to yield a robust discriminative embedding.



(a) Original Image (b) Cropped Image

Figure 4: Images in the CityFlow-ReID dataset are loosely cropped. In (a) it can be observed that almost 75% of the image contains information irrelevant to the vehicle. (b) Running a detector on the images assists in obtaining a well constrained bounding box.

demonstrates the pipeline of the proposed approach. We describe each module in the following subsections.

3.2.1 Pre-Processing

The CityFlow-ReID dataset has been gathered from real world camera feeds by applying an object detector on video frames and tracking the detected vehicles. We observed that in many cases, the images of tracked vehicles are relatively larger than the actual vehicle of interest and therefore a considerable portion of image does not contain any information about the car and might confuse the re-identification system. Consequently, we ran the Detectron [3] object detector, which implements the Mask R-CNN object detector, on all the images in the CityFlow-ReID dataset. Mask R-CNN predicts the object mask along with the bounding box which helps with tightening the predicted bounding boxes. Fig. 4 shows the importance of this step.

3.2.2 Discriminative Feature Extraction

The most important part of any re-identification system is the feature extraction module. This module must be robust since vehicles of the same make, model and color share visual similarities and there are only subtle differences that can help in differentiating them. Also, the same vehicles can have different visual appearances based on their orientation. To this end, we employ the state of the art image-based vehicle re-identification system, AAVER[6] to extract discriminative features for vehicle identities. AAVER first extracts global appearance information of the vehicle, then it estimates the orientation of the vehicle along with key-points defined in [26]. Based on the orientation of the vehicle, it selects a subset of the predicted key-points and extracts features in their vicinity. Finally, these two sets of features are combined into a single discriminative feature vector. The encoded feature vector contains information about the color, make, model and the identity of the vehicle.

According to Tang *et al.* [24], CityFlow-ReID is one of the most challenging publicly available vehicle re-identification datasets to date mostly due to illumination variations arising from climatic aberrations, occlusion, scale and quality of images. The other publicly available datasets that have similarities to CityFlow-ReID, are Veri-776 and VRIC [5].

Veri-776 dataset is composed of 49,357 images of 776 vehicle identities in a network of 20 non-overlapping cameras. The vehicle images in this dataset capture vehicles in different orientations similar to the CityFlow-ReID dataset.

VRIC dataset which is gathered from detection and tracking UA-DETRAC benchmark[27, 14], includes 60,430 images of 5,622 vehicle identities. All the images are captured by a network of 60 cameras and are of low resolution and have extreme variations in scale, and aspect ratio similar to CityFlow-ReID dataset.

Since these two datasets share similarities with CityFlow-ReID, we take advantage of them when training the AAVER model. For training the two paths of the AAVER model we used then L_2 softmax loss function suggested in [17], which can be mathematically expressed as:

$$\mathcal{L} = -\log \frac{\exp(W_y^T(\frac{\alpha \mathbf{x}}{\|\mathbf{x}\|_2}) + b_y)}{\sum_{j=1}^{N} \exp(W_j^T(\frac{\alpha \mathbf{x}}{\|\mathbf{x}\|_2}) + b_j)} \tag{1}$$

where $\mathbf{x}$ is the embedding corresponding to the input of class y, W_j and b_j are corresponding weight and bias to class label j, α is a trainable positive scalar and N is the number of classes.

3.2.3 Post-Processing

As seen in Fig. (7a), the AAVER model, results in relatively discriminative features on the challenging CityFlow-ReID dataset and separates the identities to a certain extent in the feature space. To ensure the maximum separation of identities, we apply the TPE method over the deep features generated by AAVER. TPE, a linear transformation to a lower dimensional space, couples the DCNN-based approaches and serves as a discriminative embedding step. While TPE provides the advantages of triplet learning, it can be learned relatively fast. In order to train the TPE, assume the triplet $(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n)$ in which $\mathbf{x}_a$, $\mathbf{x}_p$ and $\mathbf{x}_n$ are the n dimensional feature vectors corresponding to an anchor, its positive and negative respectively. Here we would like the similarity score of the pair $(\mathbf{x}_a, \mathbf{x}_p)$ to be larger than the similarity score of pair $(\mathbf{x}_a, \mathbf{x}_n)$ which happens with the probability of

$$p(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) = \frac{e^{\mathbf{x}_a^T \mathbf{x}_p}}{e^{\mathbf{x}_a^T \mathbf{x}_p} + e^{\mathbf{x}_a^T \mathbf{x}_n}} \tag{2}$$

Now under the linear transformation $W : \mathbb{R}^n \to \mathbb{R}^d$, TPE tries to maximize the probability that positive pairs cluster together in the d-dimensional space. TPE is learned by solving the following optimization problem:

$$\underset{W}{\text{argmin}} \sum_{(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \in \mathbb{T}} -\log\left(p(W\mathbf{x}_a, W\mathbf{x}_p, W\mathbf{x}_n)\right) \tag{3}$$

Where $\mathbb{T}$ is the set that contains all possible triplets.

After training the AAVER deep model and extracting the training set features, we train the TPE matrix over the extracted features to learn a discriminative lower dimensional representation.

The task of Nvidia AI City Challenge Track 2 considers video-based vehicle re-identification. Based on this, we initially learn discriminative features via the AAVER deep model and the TPE transformation. At the time of inference, we first group all the images in the gallery set according to the tracking information. Next, in each group of tracked images we find the similarity of the probe image to the members of the group and use the mean of the highest five similarity scores as the similarity score of the group. At this point, every group has a similarity score with respect to the query. Finally, we rank all the groups based on these similarity scores and the gallery is ranked accordingly.

4. Track 3: Vehicle Anomaly Detection

4.1. Dataset

The NVIDIA AI City Challenge provides 100 unconstrained videos in both training and testing sets for the task of vehicle anomaly detection. These videos represent real-world conditions, and varying levels of difficulty.

4.2. Video Stabilization

Many traffic cameras in this dataset have significant camera motion which causes misalignment in the detection of vehicles in consecutive frames. This negatively impacts vehicle tracking performance and causes significant tracklet fragmentation. We use an open source tool [21] for video stabilization to reduce the impact of camera motion.

4.3. Vehicle Detection

We used the Detectron object detector as the vehicle detector. Qualitative comparison between Mask R-CNN and other state-of-the art detectors reveals that the Mask R-CNN generalizes well to the domain of the NVIDIA AI City dataset and is able to detect large number of vehicles, particularly in high density scenes. For this challenge, we are interested in detecting cars, trucks, bikes, people, and traffic lights. However, Mask R-CNN has two drawbacks. First, the detector often incorrectly classifies street signs, billboards, and buildings as vehicles with a high degree of confidence. In addition, it is unable to detect several instances of anomalous vehicles due to partial occlusion, small size, or poor contrast with the background. Although our method does not address the issue of detecting small vehicles, we propose a solution to the falsely detected signs, billboards, and buildings in section 4.5.3.

4.4. Tracking

We use SORT [1] to cluster vehicles with the same identity into tracklets. SORT provides comparable accuracy to state-of-the-art methods and runs significantly faster than its deep learning based counter parts. We use a Kalman filter-based tracker rather than a deep learning based model because a majority of the tracks do not correspond to actual anomalies. Most tracklets are rejected by the first stage filtering step. Extracting deep features for all the vehicles, particularly in dense scenes, did not provide any measurable improvement in performance.

SORT provides high quality tracks in most scenarios, but it fails in two cases. First, it automatically assigns a new track identity for a given vehicle when there is partial or total occlusion. This causes track fragmentation when one ve-

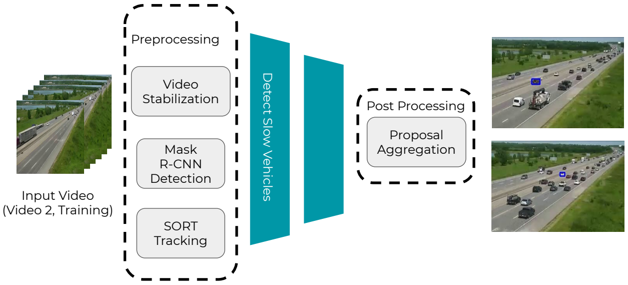


Figure 5: Proposed two-stage model for vehicle anomaly detection

hicle passes in front of another vehicle. Second, the tracker occasionally transfers a given track identity to a different vehicle when two vehicles move in close proximity to one another. This occurs when two vehicles temporarily have a large overlap in their bounding boxes and because SORT does not use discriminative features when tracking vehicles. Our method addresses both of these shortcomings in section 4.5.2.

4.5. Proposed Method

Classifying anomalous behavior requires contextual information to accurately separate stalled vehicles and accidents from normal traffic. Anomalies share several characteristics: (1) Vehicles of interest are immobile for longer than other vehicles. (2) Anomalous vehicles are significantly slower than surrounding traffic. These observations motivate the design of the proposed approach, as described in 5. We leverage these two observations about anomalous vehicles to design a two-stage approach. First, we flag vehicles that are traveling slowly. We select these initial proposals by searching for vehicles in which sequential bounding boxes have significant spatial overlap. Next, we establish a cuboid search radius around the initial proposals and compare the attributes of each anomaly proposal with those of vehicles within spatial and temporal bounds of the cuboid to refine our initial proposals.

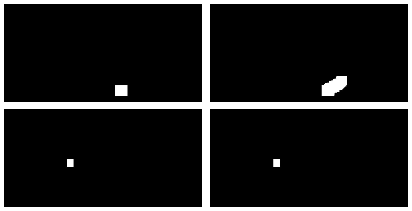


Figure 6: Stationary vehicles will generally have smaller spatial footprints when compared to moving vehicles.

4.5.1 Slow Vehicle Detection

Anomalous vehicles are generally much slower than the vehicles around them. We measure this by first splitting a given track into chunks of equal length. We then compare the Intersection over Union (IoU) of the first frame of a given chunk with the motion pattern of the total chunk. A chunk is considered anomalous if the IoU between the bounding box of the first frame and the motion pattern of the entire chunk is above a given threshold. A track is considered anomalous if a fraction of chunks is anomalous above a given threshold. For each chunk, we create two masks. As shown in Fig. 6, we set the region defined by the first frame bounding box to 1 in the first mask. Then we take the union of bounding boxes in the chunk and set the derived spatial extent to 1 in the other mask. We then compare the two masks by calculating the IoU and if the IoU is above a given threshold the corresponding vehicle is flagged as a proposal for an anomaly.

4.5.2 Proposal Merging

Inspired by [13], we consider the spatio-temporal consistency of two tracklets, and the ReID score to merge proposals. This step is essential for increasing the separation between true positives and false positives, allowing additional proposal filtering based on the length of the track.

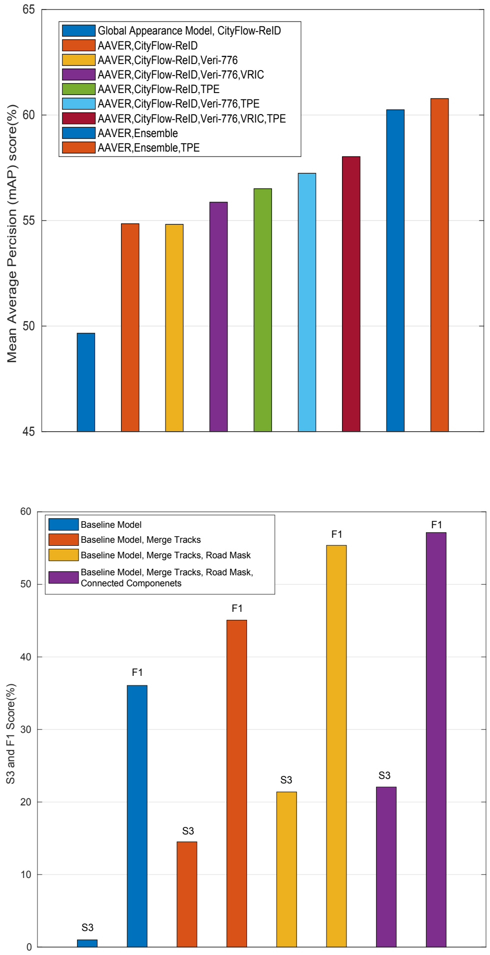
4.5.3 Context Aware Proposal Filtering

Vehicles in slow traffic will likely be detected by the first stage algorithm. However, inspection of the surrounding vehicles inform us that the exhibited behavior is normal given the behavior of the surrounding group. To accomplish this, we propose a context-based reasoning to remove proposals of parked vehicles and vehicles at traffic lights. For each proposal longer than a set threshold, we compare the pixel velocity of corresponding vehicle against its neighbors.

Lastly, after filtering all the proposals, we iterate through the original list of proposals to find tracklets that have a high spatial overlap with the final anomaly proposals and may have been improperly filtered. This additional step helps improve temporal precision.

4.5.4 Road Boundary Model

We apply the motion and location of vehicles to generate a road boundary model. Specifically, we only consider vehicles that were previously not considered as potential proposals. We plot the location of each vehicle for every frame and use this road mask to obtain vehicle's proximity to the road. We recalculate this road mask for every proposal which prevents the mask from being corrupted if the camera angle or



(b) Anomaly Detection

Figure 7: (a) Shows the progression in precision obtained after each step. It can be observed that triplet embedding and external data are crucial for the obtained results. (b) Demonstrates the enhancement that each added step brings to the overall anomaly detection system.

position is changed. In addition, this step helps to reject proposals in regions of the video with significant tearing.

4.6. Proposal Post-Processing

Finally, we merge the proposals that are within a given time interval from each other. This helps to fill in potential gaps in detections.

5. Experiments

In this section we describe the details of implementing our models for both vehicle re-identification and anomaly detection tracks.

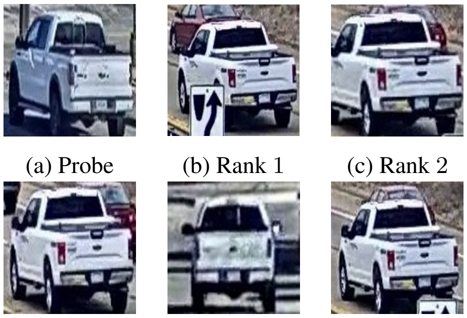
5.1. Vehicle Re-identification

The AAVER model has two paths, one for extracting global appearance features and one for extracting the orientation conditioned part appearance features. We first initialize the global appearance path of AAVER with the weights pre-trained on the CompCars dataset. Subsequently, we start training the global appearance branch on the 3,720 (333 from CityFlow-ReID, 576 from Veri-776 and 2,811 from VRIC) unique identities in their respective training sets. We used the test set of Veri-776 dataset as the evaluation set to measure the network's performance, We trained the network for 10 epochs with an initial learning rate of 10^{-4} using ADAM optimizer[7]. During training, images are randomly flipped horizontally and rotated with a random rotation angle $\theta_r \sim U(-5°, 5°)$ as means of data augmentation. After training the global appearance path of the AAVER, its weights are frozen and parameters of the part appearance path are initialized with the weights from the network trained on CompCars dataset. We trained the second path for two epochs after which the model starts to over fit to the training data. It is worth mentioning that a single global appearance branch cannot capture minute differences between vehicles of similar appearance and unsurprisingly we observe a significant improvement in performance (more than 5%) by augmenting the global appearance model with AAVER. This can also be seen in Fig. 7a

Based on the results obtained from the evaluation server, we consider three scenarios for training the second path of AAVER: Training (1) only on the CityFlow-ReID dataset, (2) on CityFlow-ReID and Veri-776 datasets, (3) CityFlow-ReID, Veri-776, VRIC datasets. Fig. 7a shows that appending different training datasets contributes towards improvement in re-identification accuracy despite the existence of domain shift across the datasets. Triplets are randomly selected and different TPE matrices are obtained for each scenario by solving the optimization problem in Eq. 3 over $10,000$ iterations. In each of the above mentioned scenarios, the overall model is capable of achieving an mAP score of greater than 56%. Therefore, in the subsequent experiment an ensemble of the above models is considered. Our motivation relies on the assumption that, depending on the dataset, the decision boundaries of each network are different and will work in conjunction by providing complementary information to each other as they are trained for the same task. This resulted in an improvement of more than 2% in mAP score and training a TPE on this ensemble improved the results further to 60.78%. Figurs 1 and 8 qualitatively shows the retrieval results of two probe images.

5.2. Anomaly Detection

Our experiments are informed by the quantitative analysis of the training set and a qualitative analysis of the test set. Each feature was crafted to generalize on the entire



(a) Probe (b) Rank 1 (c) Rank 2

(d) Rank3 (e) Rank 4 (f) Rank 5

Figure 8: Top five results of the proposed method for a given probe. We can observe that while the returned images are all visually similar and even share the same make, model and color, they do not share the same identity. Here the true match is most likely in rank 4 entry.

dataset. This approach helped us improve our F1 score as seen in Fig. 7b. Our final S3 score is 22.07%. This score is comprised of the product of F1 score and NRMSE score. Our F1 score is 57.14% and our NRMSE is 0.6137.

Our experiments are limited by the quality of detections, length of each vehicle tracklet, and the number of hyperparameters. Each external module was used off-the-shelf, and was not fine-tuned on the specific domain. However, application specific detection and tracking would have increased the effectiveness of our context aware modules. While our proposed model localizes considerable number of anomalous vehicles in the track 3 dataset, it suffers from false positives. As seen in Fig. 9, our module incorrectly classifies vehicles in parking lots as anomalous if no other vehicles in the parking lot are detected. Our search space is arbitrarily large and extends onto the road, giving improper context for that particular proposal. Since these false positives are stationary, they obscure the temporal bounds of a true anomaly, adversly affecting our performance on the S3 evaluation metric.

6. Conclusion

In this work we approached two tracks of the Nvidia AI City Challenge 2019, namely Vehicle Re-identification and Anomaly Detection.

For the Vehicle Re-Identification track we presented a robust vehicle re-identification system that relies on the highly discriminative features extracted by the AAVER model and further increases the accuracy of features while reducing their dimensionality using the method of triplet probabilistic embedding. Moreover, we make use of external datasets with similar characteristics and the ensemble of feature extractors which considerably improved the accuracy of the overall Re-ID model.

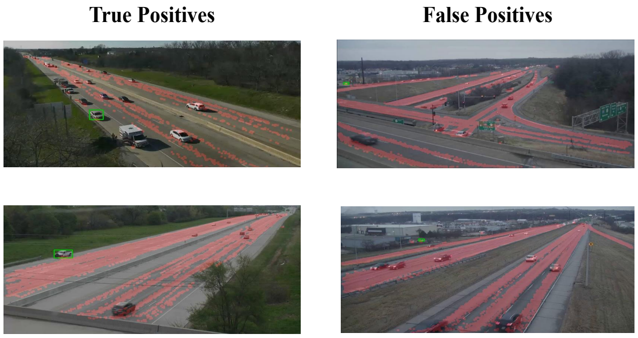


Figure 9: Despite accurately localizing many anomalies, the temporal bounds of our predictions are obfuscated by false positive detections

For the anomaly detection track we proposed a non-deep learning approach that considers local context in order to classify a vehicle. We developed this algorithm by observing trends in anomaly incidents and the behavior of surrounding vehicles.

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] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *CoRR*, abs/1602.00763, 2016.

[2] C. Cui, N. Sang, C. Gao, and L. Zou. Vehicle reidentification by fusing multiple deep neural networks. In *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2017.

[3] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] Aytac Kanaci, Xiatian Zhu, and Shaogang Gong. Vehicle re-identification in context. In *Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, September 10-12, 2018, Proceedings*, 2018.

[6] 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. *arXiv preprint arXiv:1905.03397*, 2019.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2167–2175, 2016.

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

[10] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European Conference on Computer Vision*, pages 869–884. Springer, 2016.

[11] X. Liu, W. Liu, T. Mei, and H. Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018.

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

[13] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. In *The 2018 DAVIS Challenge on Video Object Segmentation-CVPR Workshops*, 2018.

[14] Siwei Lyu, Ming-Ching Chang, Dawei Du, Longyin Wen, Honggang Qi, Yuezun Li, Yi Wei, Lipeng Ke, Tao Hu, Marco Del Coco, et al. Ua-detrac 2017: Report of avss2017 & iwt4s challenge on advanced traffic monitoring. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–7. IEEE, 2017.

[15] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. Angry crowds: Detecting violent events in videos. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 3–18, Cham, 2016. Springer International Publishing.

[16] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. Analyzing tracklets for the detection of abnormal crowd behavior. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 148–155, Jan 2015.

[17] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[18] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, April 2017.

[19] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2016.

[20] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *CoRR*, abs/1703.05921, 2017.

[21] Adam Spannbauer. A python package to stabilize videos using opencv. `https://github.com/AdamSpannbauer/python_video_stab`, 2018.

[22] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *CoRR*, abs/1801.04264, 2018.

[23] Y. Tang, D. Wu, Z. Jin, W. Zou, and X. Li. Multi-modal metric learning for vehicle re-identification in traffic surveillance environment. In *IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258, 2017.

[24] 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 *CVPR 2019: IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[25] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 379–387, 2017.

[26] 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 *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint arXiv:1511.04136*, 2015.

[28] F. Wu, S. Yan, J. S. Smith, and B. Zhang. Joint semi-supervised learning and re-ranking for vehicle re-identification. In *IEEE Conference on Pattern Recognition (ICPR)*, 2018.

[29] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3973–3981, 2015.

[30] Y. Zhou and L. Shao. Aware attentive multi-view inference for vehicle re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6489–6498, 2018.