Image-based Vehicle Re-identification Model with Adaptive Attention Modules and Metadata Re-ranking

Quang Truong

Abstract

Vehicle re-identification is a challenging task due to intra-class variability and inter-class similarity across non- overlapping cameras. To tackle these problems, recently proposed methods require additional annotation to extract more features for false positive image exclusion. In this paper, we propose a model powered by adaptive attention modules that requires fewer label annotations but still out-performs the previous models. We also include a re-ranking method that takes account of the importance of metadata feature embeddings in our paper. The proposed method is evaluated on CVPR AI City Challenge 2020 dataset and achieves mAP of 37.25% in Track 2.

1. Introduction

In recent years, computer vision has achieved milestones across its sub-fields thanks to the continuing development of Convolutional Neural Network (CNN). However, among sub-fields of computer vision, object re-identification has gained attention due to several technical difficulties. The first challenge with object re-identification is intra-class variability. Because of illumination conditions, obstacles, and occlusions, an object may appear different across non-overlapping cameras. The second challenge is the inter-class similarity. Two objects may share similar looks, such as identical twins or cars from the same manufacturing process. Unlike image classification, which simply classifies images based on visual contents, object re-identification demands a robust system to respond to local features and global features. Local features involve differentiating two objects with similar viewpoints. In contrast, global features involve clustering images that belong to the same objects, regardless of viewpoints. Re-identification systems also have to possess a good generalization ability to deal with unseen features due to plenty of object variations.

In initial studies, the majority of the research on re-identification was focused on person re-identification. Vehicle re-identification has successfully adopted the contributions of that previous research despite the difference of domains [6, 11, 15, 21, 19, 14, 13, 2, 16, 23, 17, 7, 9]. However, the majority of these projects adopt the pre-trained ImageNet classification-specific models and perform transfer learning for the vehicle re-identification task. Our proposed method focuses on GLAMOR, a model designed for re-identification proposed by Suprem *et al.* [19], which proves that training from scratch with a smaller dataset (36, 935 real images and 192, 150 synthetic images versus 14M images of ImageNet) does not necessarily result in poorer performance. In fact, GLAMOR outperforms ResNet50 baseline with 7.9% mAP improvement [21]. We also propose a slight modification to k-reciprocal encoding re-ranking [30] so that it includes the metadata attributes during the re-ranking process. The remainder of the paper is structured as follows: Section 2 reviews the related work, Section 3 illustrates our proposed approach, Section 4 focuses on our experiment, and Section 5 draws a conclusion and discusses potential rooms for improvement to study the re-identification problem.

2. Related Work

Re-identification problems have been a challenging task in computer vision. Unlike image classification, where images are required to be classified into classes, re-identification is designed to identify a probe image in an image gallery. While image classification achieves successful results [20, 22, 25], thanks to large popular datasets such as COCO [12] or ImageNet [3][10], re-identification is yet to have sufficiently large datasets to train a model. DukeMTMC [4] and Market-1501 [29] are datasets specifically for person re-identification, while Veri-776 [14] and VehicleID [13] are for vehicle re-identification. These datasets share a common disadvantage, which is the lack of images per identity. Intra-class variability and inter-class similarity are also common problems in re-identification due to diverse backgrounds or similar looks.

Novel approaches to overcome the above disadvantages have been proposed recently. Hermans *et al*. prove that triplet loss [24, 18, 2] is suitable for re-identification task since it optimizes the embedding space so that images with the same identity are closer to each other compared to those with different identities [6]. Hermans *et al.* also propose the Batch Hard technique to select the hardest negative samples within a batch, minimizing the intra-class variability of an identity [6, 2, 11].
Besides data mining techniques and alternative loss functions, there have been several efforts to implement new models designed for re-identification [19, 13, 6, 23, 9]. Specifically, Suprem *et al*. focus on using attention-based regularizers [19, 9] to extract more global and local features and ensure low sparsity of activations. Wang *et al*. utilize 20 key point locations to extract local features based on orientation thanks to attention mechanism, and then fuse the extracted features with global features for orientation- invariant feature embedding [23].

Re-ranking is also an important post-processing method that is worth considering in re-identification. Zhong *et al*. propose a re-ranking method that encodes the k-reciprocal nearest neighbors of a probe image to calculate k-reciprocal feature [30]. The Jaccard distance is then re-calculated and combined with the original distance to get the final distance. Khorramshahi *et al*. utilize triplet probabilistic embedding [17] proposed by Sankaranarayanan *et al*. to create similarity score for re-ranking task [9]. Huang *et al*. propose meta- data distance, which uses classification confidence and confusion distance. Metadata distance is then combined with the original distance to get the final distance [7].

3. Proposed Approach

3.1. System Overview

The overview of our system can be categorized into three main stages: pre-processing, deep embedding computing, and post-processing. The system is described in Figure 1. Pre-processing is necessary since the bounding boxes of the provided dataset are loosely cropped. The loosely cropped images contain unnecessary information, which hinders the performance of our model.

The deep metric embedding module is a combination of GLAMOR [19] and Counter GLAMOR, and is trained on the provided dataset. The output of the module is a W × H distance matrix where W represents the images in query and H represents the images in the gallery. Additional classifiers are also trained on the provided dataset to extract metadata attributes for further post-processing.

Post-processing is essential in re-identification since it removes false-positive images at the top. Illumination conditions, vehicle poses, and other various factors affect the outputs negatively. Figure 2 shows an example of two images with close embedding distance due to similarities in brightness, pose, color, and occlusion.

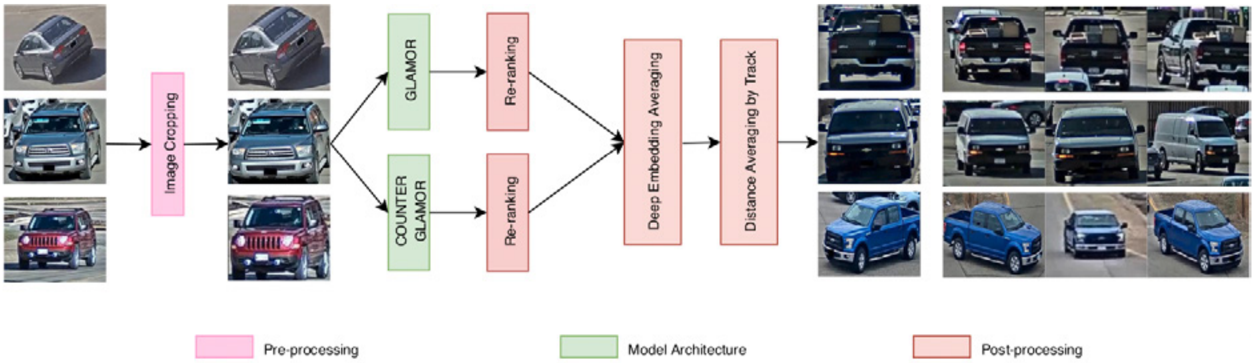


Figure 1. System Overview.

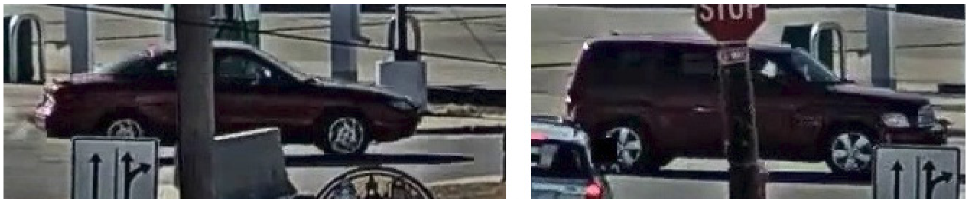


Figure 2. An example of two images with close embedding distance due to similar brightness, pose, color, and occlusion.

3.2. Pre-processing

3.2.1 Detectron2

We adopt pretrained Detectron2 [27] on MS COCO dataset [12] to first detect a vehicle in an image and then to crop the bounding box out of the image. Detectron2 is a Facebook platform for object detection and segmentation that implements state-of-the-art object detection algorithms, including Mask R-CNN[5]. We perform image cropping on training, query, and test sets; we then use the cropped images for training as well as evaluating models.

3.2.2 Image Labeling for Vehicle Attribute Extractor

As shown in Figure 2, car type does not match. Even though they have close embedding distance, the embedding distance is mostly affected by noise features. Therefore, vehicle metadata attributes should be extracted to eliminate undesired features such as obstacles in the background.
We adopt pre-trained ResNeXt101[28, 1] on ImageNet [10] for rapid convergence. We train ResNext101 to classify color and type.

The given color labels and type labels do not reflect the training set. For example, the training set does not contain any orange cars. The number of cars per category is also unevenly distributed; there is a lack of RV or bus images. Therefore, we cluster types based on their common visual attributes. For types, we suggest having six categories: small vehicle (sedan, hatchback, estate, and sports car), big vehicle (SUV and MPV), van, pickup, truck, and long car (bus and RV). For color, we exclude orange, pink, and purple.

The query set and test set, however, do contain the excluded categories. Moreover, there are different cameras in the query and test sets. The training set is collected from thirty-six cameras while the query and test sets are collected from twenty-three cameras, so performing prediction on the query and test sets will eventually result in incorrect classification. Therefore, we extract the features before the last fully-connected layer and calculate the Euclidean distance between the query set and the test set for the re-ranking process.

3.3. Deep Embedding Computing

We adopt GLAMOR, an end-to-end ResNet50- backboned re-identification model powered by attention mechanism, proposed by Suprem *et al.* [19]. GLAMOR introduces two modules. The Global Attention Module reduces sparsity and enhances feature extraction. In the meantime, the Local Attention Module extracts unsupervised part-based features. Unlike the original model, we have modified the model slightly to increase the performance. Instead of using the original Local Attention Module, we use the Convolutional Block Attention Module (CBAM) as our local feature extractor because CBAM focuses on two principal dimensions: spatial and channel [25, 26]. As a feature adaptive refinement module, CBAM effectively learns where to emphasize or suppress the information that will be passed forward to the later convolutional blocks. The detailed architecture of GLAMOR is represented in Figure 3.

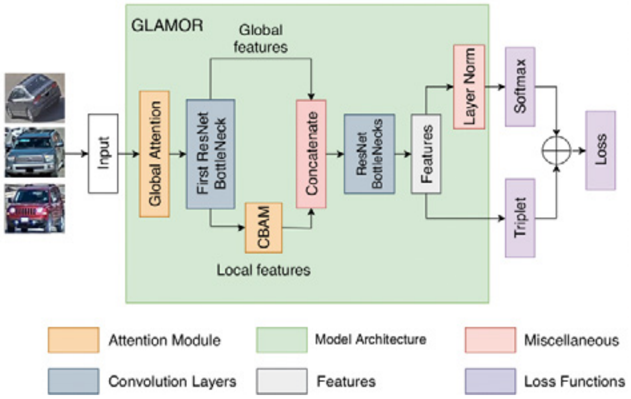


Figure 3. Architecture of GLAMOR

We also realize that there is a loss of information in the current GLAMOR implementation at the concatenation step. Suprem *et al*. apply a channel-wise mask to combine global features and local features [19]. However, only half of each is fed forward to later convolutional blocks. The sum of global features F_G and local features F_L, where $F_G, F_L \in \mathbb{R}^{H \times W \times C}$, is calculated as follows:

$$F = M_G \odot F_G + M_L \odot F_L, \qquad (1)$$

where $M_G, M_L \in \mathbb{R}^C, M_L = \bar{M_G}$, and for each $m_i \in M_G$, $m_i = 0 \ \forall i < \lfloor \frac{C}{2} \rfloor$ and $m_i = 1 \ \forall i \geq \lfloor \frac{C}{2} \rfloor$. Therefore, we propose another concatenation formula to counter the loss of information in Equation (1) by swapping the mask position:

$$F = M_L \odot F_G + M_G \odot F_L. \qquad (2)$$

The concatenation formula in Equation (2) is used for another GLAMOR. The distance embedding matrix of two GLAMORs is then averaged for the final result. The proposed method significantly increases the accuracy due to generalization and balancing effects.

The two models are trained separately on both synthetic data and training data. Training models on a synthetic dataset helps models converge faster than training on the real dataset alone. Our models converge in $20 - 30$ epochs, while a pre-trained ResNet50 baseline model converges after 60 epochs [21].

Our metric learning method is a combination of batch hard triplet loss [24, 18] and softmax loss with label smoothing [20]. The reason for this combination is that triplet loss is used for learning embeddings whereas softmax loss interprets probability distributions of a list of potential outcomes. The combination loss is

$$\mathcal{L}_{\text{TriSoft}} = \lambda_{\text{Triplet}} \cdot \mathcal{L}_{\text{Triplet}} + \lambda_{\text{Softmax}} \cdot \mathcal{L}_{\text{Softmax}}, \qquad (3)$$

where λTriplet and λSoftmax are hyperparameters that can be fine-tuned. The revised triplet loss proposed by FaceNet [18] is

$$\mathcal{L}_{\text{Triplet}} = \sum_{\substack{a,p,n \\ y_a = y_p \neq y_n}} [m + D_{a,p} - D_{a,n}]_+, \qquad (4)$$

where *a, p, n* are anchor, positive, and negative samples of a triplet, $D_{a,p}$ and $D_{a,n}$ are the distance from an anchor sample to a positive sample and to a negative sample, and *m* is the margin constraint. The softmax with label smoothing proposed by Szegedy *et al.* [20] is

$$\mathcal{L}_{\text{Softmax}} = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases}$$

and:

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon & if \quad i = y \\ \varepsilon / N & \text{otherwise,} \end{cases}$$

where y is the ground truth ID label, *pi* is the ID prediction logits of class *i*, *N* is the number of IDs in the dataset, and *ε* is a hyperparameter to reduce over-confidence of classifiers.

3.4 Post-processing

3.4.1 Re-ranking

We adopt the re-ranking with *k*-reciprocal encoding method [30] proposed by Zhong *et al.* and modify the formula to include Euclidean distance embedding of metadata attributes. Given a probe image *p* and a gallery image g_i ∈ G where G is gallery set, the revised original distance matrix is

$$d'(p, g_i) = d(p, g_i) + \sum \gamma_j \cdot D_j(p, g_i), \qquad (5)$$

where $d(p,g_i)$ is the original distance between *p* and g_i, *γj* is the hyperparameter of feature *j* for fine-tuning, and $D_j(p, g_i)$ is the metadata distance between *p* and g_i of feature *j*. We then generate the *k*-reciprocal nearest neighbor set *R* and re-calculate the pairwise distance between the probe image *p* and the gallery image g_i using Jaccard distance and a more robust k-reciprocal nearest neighbor set *R**:

$$d_J(p, g_i) = 1 - \frac{|\mathcal{R}^*(p,k) \cap \mathcal{R}^*(g_i,k)|}{|\mathcal{R}^*(p,k) \cup \mathcal{R}^*(g_i,k)|}. \qquad (6)$$

The final distance embedding is:

$$d^*(p, g_i) = (1-\lambda)d_J(p, g_i) + \lambda d'(p, g_i). \qquad (7)$$

3.4.2 Distance Averaging by Track

Given the test track for each test image, we calculate the average distance between a probe image *p* and a track. Then, we replace the distance between the probe image and each image in that track with the calculated average distance. The problem becomes finding tracks that have the most similar car to the probe image instead of finding individual im- ages. The method increases mAP since the top results will be populated with correct images from the same track for uncomplicated cases.

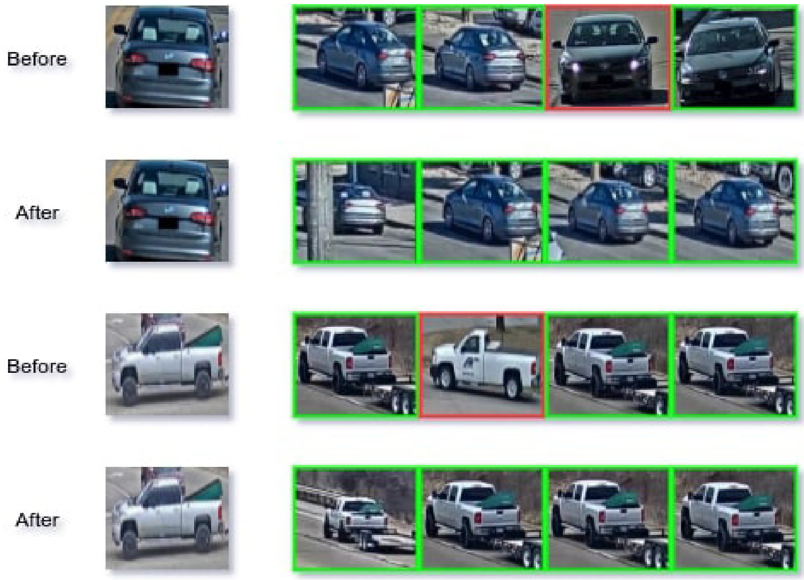


Figure 4. The effects of the re-ranking method.

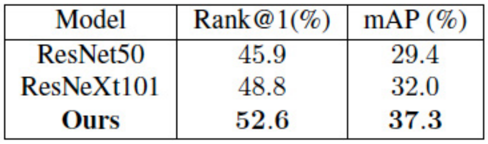
4. Experiment

Based on [8], we have enough resources for building our models with the provided utilities. After being cropped with Detectron2[27], the images are resized to 128 × 256 for training GLAMOR models [19] and 224 × 224 for training ResNeXt101 model [28]. Image size may largely affect the re-identification results according to [15]; therefore, we choose 128 × 256 as our image size because vehicle images tend to have the width larger than the height. The default image size of the pre-trained

ResNeXt101 is 224 × 224, so we keep it in order to transfer learning efficiently. The im- ages are then augmented with flipping and cropping techniques, color jitter, color augmentation[10], and random erasing [31].

The GLAMOR models are pre-trained with the synthetic data for around 25 epochs with an initial learning rate of 0.0002, learning rate decay of 0.2 for every 10 epochs, a margin of 0.3, and the 10 : 1 ratio between triplet loss and softmax loss. After that, we feed the transformed images above to the GLAMOR models for the re-identification task with similar parameters. The models converge quickly in around 25 epochs thanks to the pre-trained weights.

We repeat the same procedure with ResNext101 but with pre-trained weights from ImageNet [10, 1], instead of the synthetic data. After training for 55 epochs for re- identification task with an initial learning rate of 0.0003, learning rate decay of 0.3 for every 20 epochs, and the same margin and loss ratio, we keep that weight to train the ResNext101 models further to classify type and color. For the classification task, we train the models using soft- max loss only with the learning rate of 0.002 and learning rate decay of 0.5 for every 20 epochs.



Model	Rank@1(%)	mAP (%)
ResNet50	45.9	29.4
ResNeXt101	48.8	32.0
Ours	**52.6**	**37.3**

Table 1. Comparison with base line models.

Even though we have weights of two different models GLAMOR and ResNeXt101 for re-identification tasks, we find that GLAMOR outperforms ResNext101. There- fore, we decide to use only GLAMOR models for the re- identification task. On the other hand, since ResNext101 is a state-of-the-art image classification model, we use it as our metadata attribute extractor.

Our proposed approach achieves mAP of 37.25% in Track 2 of the AI City Challenge 2020. Table 1 compares our result with two different base line results provided in [21].

5. Conclusion

In this paper, we introduce an attention-driven re-identification method based on GLAMOR [19]. We also incorporate metadata attribute embedding in the re-ranking process, which boosts the performance of the model. In addition, several techniques in pre-processing and post-processing are adopted to enhance the results. Below are topics that should be further studied in order to improve our system:

- Image super-resolution for pre-processing.
- GAN-based models in vehicle re-identification.
- View-aware feature extraction.
- Intensive hyperparameter tuning.

References

[1] Remi Cadene. Pretrained models for Pytorch. https://github.com/Cadene/pretrained-models.pytorch, 2019.

[2] G. Chen, T. Zhang, J. Lu, and J. Zhou. Deep meta metric learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9546–9555, 2019. http://ivg.au.tsinghua.edu.cn/people/Guangyi_Chen/1040_camera_ready_final.pdf

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei- Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf

[4] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke. DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017. 1 https://openaccess.thecvf.com/content_cvpr_2017_workshops/w17/papers/Gou_DukeMTMC4ReID_A_Large-Scale_CVPR_2017_paper.pdf

[5]. K. He, G. Gkioxari, P. Dolla´r, and R. B. Girshick. Mask R-CNN. CoRR, abs/1703.06870, 2017. 2 https://arxiv.org/abs/1703.06870

[6] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification, 2017. 1, 2 https://arxiv.org/abs/1703.07737

[7] T. Huang, J. Cai, H. Yang, H. Hsu, and J. Hwang. Multi- view vehicle re-identification using temporal attention model and metadata re-ranking. In The IEEE Conference on Com- puter Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1, 2 https://openaccess.thecvf.com/content_CVPRW_2019/papers/AI%20City/Huang_Multi-View_Vehicle_Re-Identification_using_Temporal_Attention_Model_and_Metadata_Re-ranking_CVPRW_2019_paper.pdf

[8] Jakel21. Vehicle ReID baseline. https://github.com/Jakel21/vehicle-ReID-baseline, 2019. 4

[9] P. Khorramshahi, N. Peri, A. Kumar, A.l Shah, and R. Chellappa. Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1, 2 https://openaccess.thecvf.com/content_CVPRW_2019/papers/AI%20City/Khorramshahi_Attention_Driven_Vehicle_Re-identification_and_Unsupervised_Anomaly_Detection_for_Traffic_CVPRW_2019_paper.pdf

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural net- works. Commun. ACM, 60(6):84–90, May 2017. 1, 2, 4 https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[11] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram. Vehicle re-identification: an efficient baseline using triplet embedding, 2019. 1, 2. https://arxiv.org/pdf/1901.01015.pdf

[12] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dolla´r. Microsoft COCO: Common objects in context, 2014. 1, 2. https://arxiv.org/pdf/1405.0312.pdf

[13] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2167–2175, 2016. 1, 2. https://openaccess.thecvf.com/content_cvpr_2016/papers/Liu_Deep_Relative_Distance_CVPR_2016_paper.pdf

[14] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re- identification in urban surveillance videos. In 2016 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, 2016. 1. DOI: https://doi.org/10.1109/ICME.2016.7553002

[15] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification, 2019. 1, 4. https://arxiv.org/pdf/1903.07071.pdf

[16] K. Nguyen, T. Hoang, M. Tran, T. Le, N. Bui, T. Do, V. Vo- Ho, Q. Luong, M. Tran, T. Nguyen, T. Truong, V. Nguyen, and M. Do. Vehicle re-identification with learned representation and spatial verification and abnormality detection with multi-adaptive vehicle detectors for traffic video analysis. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1 https://openaccess.thecvf.com/content_CVPRW_2019/html/AI_City/Nguyen_Vehicle_Re-identification_with_Learned_Representation_and_Spatial_Verification_and_Abnormality_CVPRW_2019_paper.html

[17] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chel- lappa. Triplet probabilistic embedding for face verification and clustering. 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), Sep 2016. 1, 2. https://doi.org/10.1109/BTAS.2016.7791205

[18] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A uni- fied embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recogni- tion (CVPR), Jun 2015. 2, 3, 4. https://doi.org/10.1109/CVPR.2015.7298682

[19] A. Suprem and C. Pu. Looking GLAMORous: Vehicle re-id in heterogeneous cameras networks with global and local attention, 2020. 1, 2, 3, 4, 5. https://arxiv.org/pdf/2002.02256.pdf

[20] C.Szegedy, V.Vanhoucke, S.Ioffe, J.Shlens,and Z.Wojna. Rethinking the Inception architecture for computer vision, 2015. 1, 3, 4. https://arxiv.org/pdf/1512.00567.pdf

[21] Z. Tang, M. Naphade, M. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J. Hwang. CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, 2019. 1, 3, 5. https://www.groundai.com/project/cityflow-a-city-scale-benchmark-for-multi-target-multi-camera-vehicle-tracking-and-re-identification/1

[22] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification, 2017. 1. https://arxiv.org/pdf/1704.06904.pdf

[23] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 379–387, 2017. 1, 2. https://openaccess.thecvf.com/content_ICCV_2017/papers/Wang_Orientation_Invariant_Feature_ICCV_2017_paper.pdf

[24] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009. 2, 3. https://jmlr.csail.mit.edu/papers/volume10/weinberger09a/weinberger09a.pdf

[25] S. Woo, J. Park, J. Lee, and I. S. Kweon. CBAM: Convolutional block attention module, 2018. 1, 3. https://arxiv.org/pdf/1807.06521.pdf

[26] S. Woo, J. Park, J. Lee, and I. S. Kweon. Official PyTorch code for ”BAM: Bottleneck Attention Module (BMVC2018)” and ”CBAM: Convolutional Block Atten- tion Module (ECCV2018)”. https://github.com/ Jongchan/attention-module, 2019. 3.

[27] Y. Wu, A. Kirillov, F. Massa, W. Lo, and R. Girshick. Detectron2. https://github.com/ facebookresearch/detectron2, 2019. 2, 4

[28] S. Xie, R. Girshick, P. Dolla ́r, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks, 2016. 2, 4. https://arxiv.org/pdf/1611.05431.pdf

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 1. https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Zheng_Scalable_Person_Re-Identification_ICCV_2015_paper.pdf

[30] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding, 2017. 1, 2, 4. https://arxiv.org/pdf/1701.08398.pdf

[31] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation, 2017. 4. https://arxiv.org/pdf/1708.04896.pdf