Unsupervised Meta Generative Object Re-identification

Report submitted for the Master Thesis Project

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE

by

BIMAN GIRI(MT20006)

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY (IIITD)
NEW DELHI 2022

YEAR OF SUBMISSION
2022

CERTIFICATE

It is certified that the work contained in the thesis titled **Unsupervised Meta Generative Object Re-identification** by **Biman Giri(MT20006)** has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of **Master of Technology**.

Supervisor
Dr A V Subramanyam
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology (IIITD)
New Delhi 2022

DECLARATION BY THE CANDIDATES

I, **Biman Giri** , certify that worked on a project entitled **Unsupervised Meta Generative Object Re-identification**, in partial fulfillment of the requirements for the award of Degree of **Master of Technology** in the Department of Computer Science and Engineering at **Indraprastha Institute of Information Technology, Delhi(IIITD)** in an authentic record of our own work under the supervision of **Dr. A V Subramanyam** from **May 2021** to **May 2022**.

(Biman Giri)

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the students is correct to the best of my/our knowledge.

Supervisor
Dr A V Subramanyam
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology (IIITD)
New Delhi 2022

Signature of Head of Department

Abstract

Unsupervised domain adaptation severely suffers from huge domain gap in fine grained recognition tasks such as vehicle re-identification (re-id). Existing works either focus on fully unsupervised methods using tracklet or classical clustering based progressive learning techniques. However, these methods do not leverage the available large datasets which are labelled. Though this can be done using existing unsupervised domain adaptation techniques, the fine grained nature of vehicle re-id precludes from obtaining a sound performance. To this end, we propose a joint learning framework which disentangles ID and non-ID features and enforce the adaptation module to focus on the ID features only. Our model performs the following three steps; (i) the model encodes cross domain images into shared ID space and domain specific non-ID space, (ii) adaptation is performed using adversarial domain alignment and pseudo-label generation, and (iii) meta learning is applied to obtain better generalization. We perform experiments on AI CITY, VRIC and VeRI-776, and compare against various unsupervised techniques to show the efficacy of our model.

Keywords: Vehcile Re-identification, Feature desentangling, Pseudo Labeling, Domain Adaptation , Meta Learning

Acknowledgments

We are highly grateful to **Dr. A V Subramanyam**, Associate Professor at, Indraprastha Institute of Information Technology (IIIT-Delhi), India, for providing an opportunity to carry out the project on **Unsupervised Meta Generative Object re-identification**. His immense knowledge and guidance helped us all the time of the project.

Biman Giri(MT20006)

Contents

Chapter 1

Introduction

Vehicle Re-identification is the process of re-identifying the vehicle images captured from the non-overlapping camera views given the query image. This is an essential part of the modern traffic management systems. These systems generally consists of various functionalities which includes detecting traffic elements, tracking the elements, counting the total number of vehicles in intersections, estimating the pose of the vehicles [32].
Supervised methods have achieved a significant performance in vehicle Re-identification [19, 24, 27–29]. However, when the labels are not available, the models trained on other dataset show poor performance. This is because of the domain gap between training and testing data. This domain gap refers to change of background, viewpoints and illuminations etc. To address this problem, various unsupervised approaches have been proposed [8, 12, 23, 31]. However, these methods do not fully utilize the availability of large scale annotated datasets. Thus, our thesis focuses on unsupervised domain adaptation which can efficiently utlilize the available annotated datasets.

In contrast to to traditional unsupervised domain adaptation problems such as image classification or image segmentation, vehicle re-identification problem is more challenging due to its fine-grained and open world recognition nature. Our proposed model is illustrated in Figure 1.1

Our main contributions are as follows. First, we propose a joint learning framework which disentangles the id and style features so that the adaptation can be done only in the shared id-space. Second, we introduce the cross domain cyclic consistency image generation to achieve the desired style transfer between the source and target domain. Third, we apply meta learning to achieve batter generalization and perform extensive ablation study to prove the robustness and effectiveness our model.

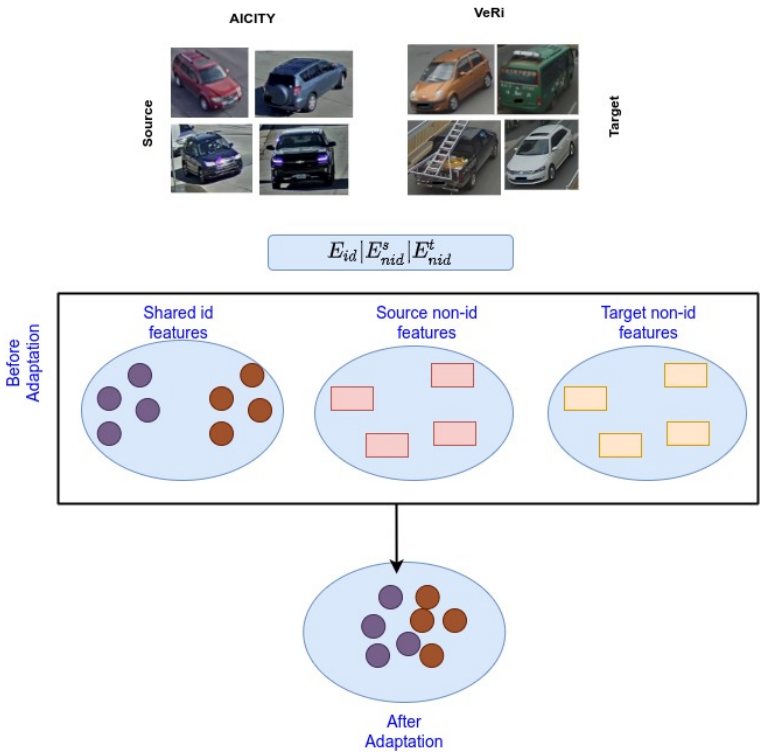


Figure 1.1: Illustration.

Chapter 2

Related Work

2.1 Disentangling

Disentangling is the process of independent features representation. In a conventional deep learning approach, disentangling is done using auto-encoder with the adversarial training [21] within a set of labelled information. InfoGAN [2] and $\beta - GAN$ [9] introduced to learn more interpretable features.

2.2 Unsupervised Domain Adaption

Unsupervised domain adaptation is gaining popularity in problems like image classification, image segmentation and object detection and re-identification. Unsupervised domain adaption is basically done at input level or feature level. At input level, the adaptation is mostly done by transferring the style between the domains [11,14]. Adaptation at features level is mostly done by minimizing the distance between source and target domain feature space using correlation [25], maximum mean discrepancy (MMD) [20] and many other methods. But in most of the works, they treat both adaptation methods separately. In our case, we try to solve both the adaptation method simultaneously.

2.3 Vehicle Re-identification

Vehicle re-id has witnessed a good number of supervised work [15,17,32]. As the labels are hard to obtain, unsupervised methods have been proposed [4,6,16,26].

Chapter 3

Component of architecture

3.1 Dataset Details

Here we have used three datasets. Our source dataset is AICITY [22] and target dataset is VeRi-776 [18] which is unlabelled. We also used another dataset called VERIC [13].
AICITY has total of 880 vehicles, out of which they used 440 vehicles for training and remaining 440 vehicles for testing. There total of 85058 vehicles in dataset, out of which 31238 images used for testing and remaining 52717 used for training. Entire dataset is prepared from images obtained from 40 different cameras.
VERIC has 2811 different vehicles with 54808 images for training and remaining 2811 identities with 5622 images for testing. The images are captured from 20 different cameras.
Veri-776 has total 776 different vehicles, out of which 200 vehicles used testing and remaining 576 vehicles used for training. There are 11579 images in the test set and 37778 images for training. Entire dataset is prepared by capturing from 20 different cameras.

3.2 Id-encoder

In our architecure, we have used resnet-50 [7] pretrained on the ImageNet [3] dataset. We further pre-train it on source dataset. Id-encoder is shared between source and target domain. Id-encoder is used to extract all the id related features like appearance and other semantics. After extracting all the features for a given image we apply multi-class cross entropy loss to enforce the id encoder to extract only id related features Figure 3.1.

3.3 DBSCAN

As the target domain does not have any labels, we apply DBSCAN to obtain pseudo-labels Figure 3.2

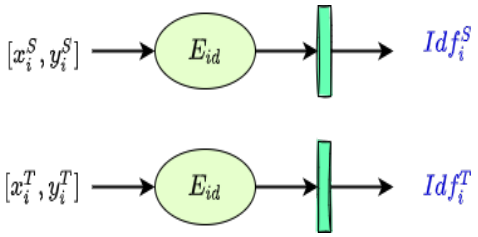


Figure 3.1: Given X^s source images and X^T target image set which is fed into E_{id} id encoder to source id features Idf^S and target id features Idf^T to calculate Multi-class cross entropy loss to enfoce the id-encoder to extacrt only id related features

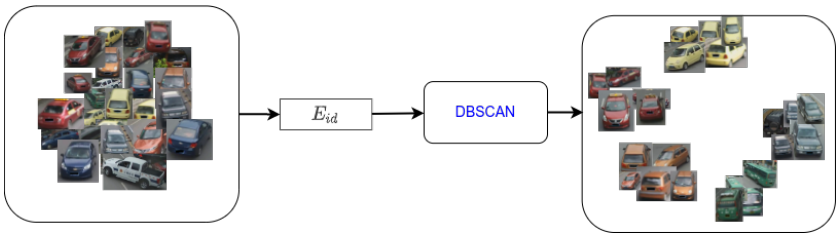


Figure 3.2: Pseudo labels generation

3.4 StyleEncoder

Style encoder extracts style features like color as shown in Figure 3.4. It is domain specific which means each domain has its own encoder. We have shown the architecture details in Figure 3.3

3.5 Domain Adversarial

Domain discriminator tries to identify the id-features membership given source and target id-features. Domain discriminator is denoted by D_{dom} and E_{id} acts as the generator Figure 3.5.

3.6 Image adversarial

At image adversarial training given any id-features from one image from any domain and style features from another domain, the job is to generate the synthetic image that cannot be discriminated from a real image from another domain. In our proposed architecture, we have two

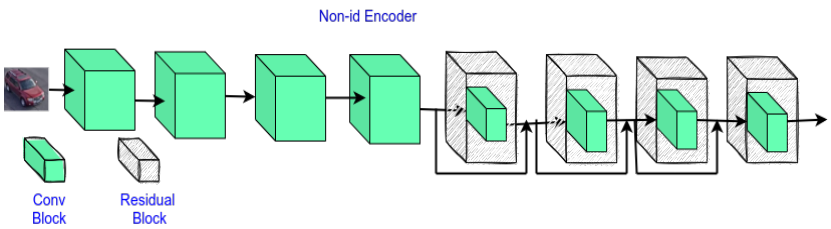


Figure 3.3: For non-id encoder we have used 4 convolution layers followed by 4 rasidual layers

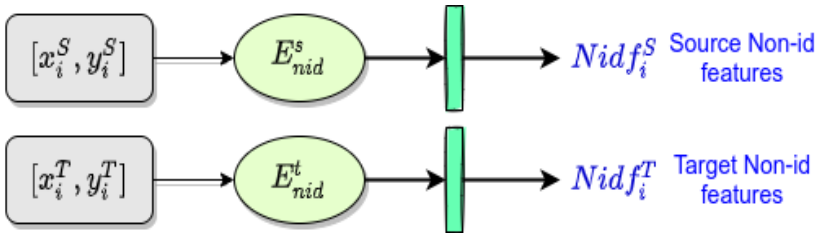


Figure 3.4: Given any X^S source images and X^T target images which extract source non-id features $Ndif^S$ and target non-id features $Ndif^T$

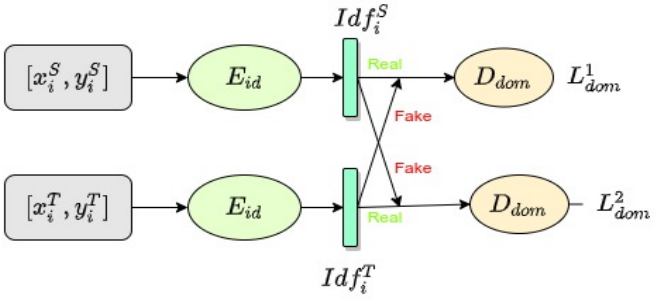


Figure 3.5: Given any X^S source images and X^T target images first E_{id} extract all source id-features and target id-features, then those are fed into D_{dom} to calculate the domain adversarial loss

generators which are domain specific and one discriminator. The architecture details of image generator is discussed in Figure 3.6 and image discriminator is PatchGAN [1].

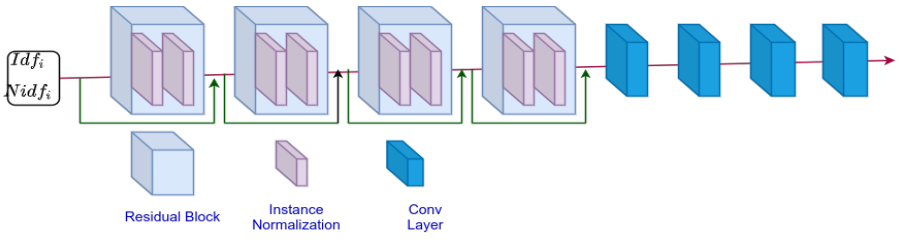


Figure 3.6: Generator is used four residual layers followed by convolution layers. Each residual layers consist of two adaptive instance normalization layers.

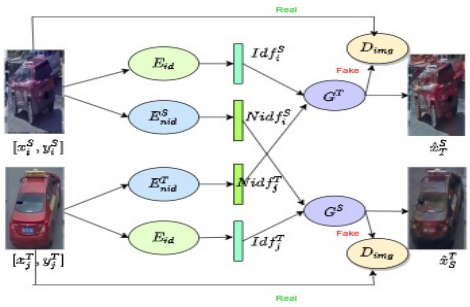


Figure 3.7: Adversarial training.

Chapter 4

Methodology

Thus, overall architecture is given in Figure 4.1.

4.1 Pseudo Label Generation

We use DBSCAN to obtain pseudo-labels

$$L_{id}^{t1} = \mathbb{E}\left[-\log p(y_j^t|x_j^t)\right] \tag{4.1}$$

where $p(y_j^t|x_j^t)$ is probability that x_j^t belong to the pseudo label y_j^t. We also calculate the identification loss on the synthetic images.

$$L_{id}^{t2} = \mathbb{E}\left[-\log p(y_j^t|x_{T_j}^S)\right] \tag{4.2}$$

$p(y_j^t|x_{T_j}^S)$ is the predicted probability that $x_{T_j}^S$ is belongs to the class y_j^t of x_j^t. We also apply multi-class cross entropy based identification loss on the source image.

$$L_{id}^{s1} = \mathbb{E}\left[-\log p(y_i^s|x_i^s)\right] \tag{4.3}$$

Where $p(y_i^s|x_i^s)$ is probability that x_j^t belong to the source label y_i^s.

$$L_{id}^{s2} = \mathbb{E}\left[-\log p(y_i^s|x_{S_i}^T)\right] \tag{4.4}$$

$p(y_i^s|x_{S_i}^T)$ is the predicted probability that $x_{S_i}^T$ is belongs to the class y_i^s of x_i^s

$$L_{id}^s = L_{id}^{s1} + L_{id}^{s2} L_{id}^t = L_{id}^{t1} + L_{id}^{t2} \tag{4.5}$$

Using pseudo labeling we enforce the shared id encoder to extract only the domain invariant features that helps the adaptation Figure 4.2.

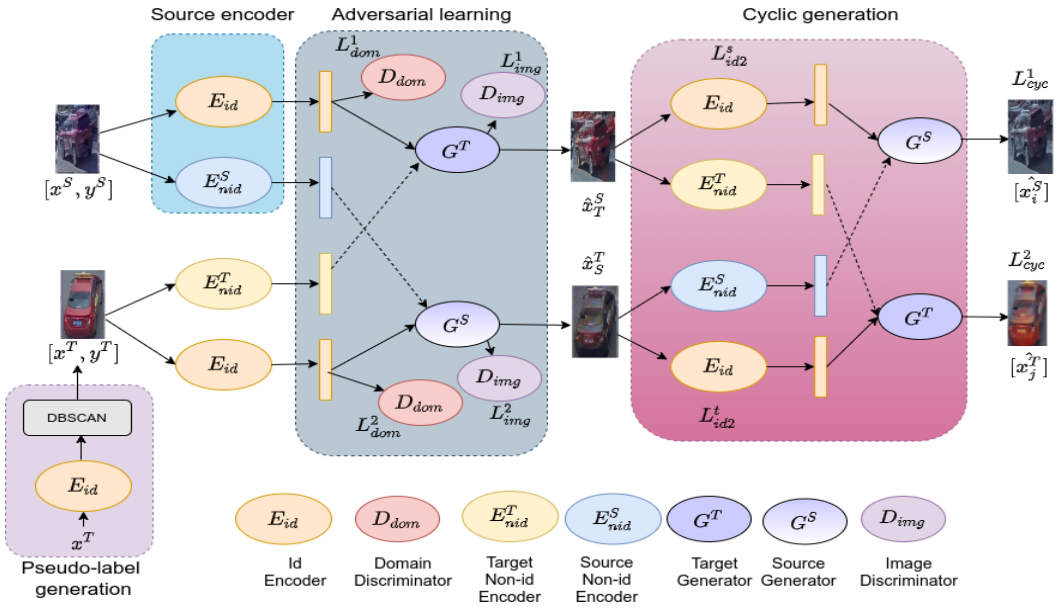


Figure 4.1: Proposed architecture

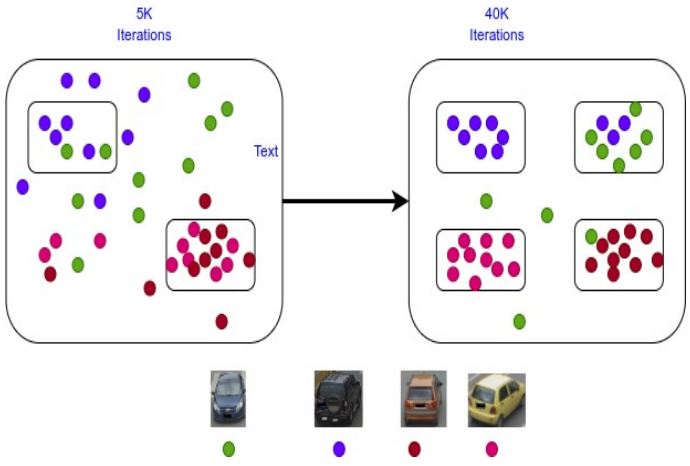


Figure 4.2: Pseudo label generation.

4.2 Domain Adversarial Module

Even though the weight of id-encoder is shared between source and target domain but it does not make sure that it will it will extract source and target id-features which is belongs to the similar distribution. So with the help of D_{dom} whose job is to differentiate the domain membership. Given x_i^s ans x_j^t the E_{id} encoder will extract source id features Idf_j^s and target id features Idf_j^t. During the domain adversarial training the E_{id} learned to extract such id features across domain so that the D_{dom} not able to differentiate the domain membership. Here we define the adversarial domain membership loss.

$$L_{dom}^1 = \mathbb{E}\left[\log D_{dom}(x_i^s) + \log\left(1 - D_{dom}(x_j^t)\right)\right] \tag{4.6}$$

$$L_{dom}^2 = \mathbb{E}\left[\log D_{dom}(x_j^t) + \log\left(1 - D_{dom}(x_i^s)\right)\right] \tag{4.7}$$

$$L_{dom} = L_{dom}^1 + L_{dom}^2 \tag{4.8}$$

4.3 Cross domain image generation

Using the cross domain image generation which force the disentangle between shared id space and domain specific non-id features space. Given source image x_i^s and target image x_j^t first using E_{id} extract id features from source Idf_i^s and from target Idf_j^t. Then extract non-id features from source Idf_i^s using E_{nid}^S and from target Idf_j^s using E_{nid}^T,we will fed Idf_i^s and $Ndif_j^t$ into G_T to generate synthesized image X_T^S which has source apperance and target structure. Similarly we will fed Idf_j^t and $Ndif_i^s$ into G^S which generate synthesized image X_S^T which has target appearance and source structure.$X_T^S = G^T(Idf_i^s, Ndif_j^t)$ and $X_S^T = G^T(Idf_j^t, Ndif_i^s)$ Now we using adversarial loss to match the distribution between synthesized image and real image.

$$L^{img} = \mathbb{E}\left[\log(D_{img}(x_i^s) + \log(1 - D_{img}(x_T^S)\right] + \mathbb{E}\left[\log(D_{img}(x_j^t) + \log(1 - D_{img}(x_S^T)\right] \tag{4.9}$$

As there are no ground truth for the synthetic image we can take the leverage of the cyclic image generation by swapping the id-features and non-id features extracted from the synthesized image.$Idf_i^s, Ndif_i^s = E_{id}(x_T^S), E_{nid}(x_S^T)$ and $Idf_j^t, Ndif_j^t = E_{id}(x_S^T), E_{nid}(x_T^S)$.

$$L_{cyc}^1 = \mathbb{E}\left[||x_i^s - G^s(Idf_i^s, Ndif_i^s)||_2\right] \tag{4.10}$$

$$L_{cyc}^2 = \mathbb{E}\left[||x_i^t - G^t(Idf_j^t, Ndif_j^t)||_2\right] \tag{4.11}$$

$$L_{cyc} = L_{cyc}^1 + L_{cyc}^2 \tag{4.12}$$

As D_{img} is shared across the domain, so it generate more realistic synthetic images also indirectly force the in-encoder to learns the domain invariant features.Here diagrammatically Figure 4.3 we are displaying how given source and target image how can we generate synthetic image, and from it how again reconstructing back to original image.

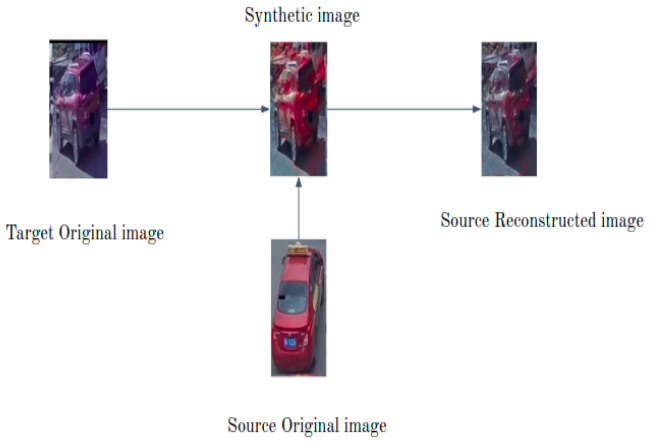


Figure 4.3: Given Target original image and source original image, we are generating synthetic image which target image appearance and source image structure(here color) and finally again reconstructing back to the original image.

4.4 Meta Learning Algorithm

For meta learning algorithm [5], we have used three datasets, where AICITY [22] ,VERIC [13] used for meta train and Veri-776 [18] for meta test.Within each epochs, we are sampling k data-points from meta train then getting meta grading for each meta train dataset. After then we samples M data points for meta update to get the final gradient.The detail algorithm is defined 1

4.5 Optimization

We train all the components shared id encoder, domain specific non-id encoder, source generator, target generator, domain discriminator, image discriminator to optimize the total objective function which is weighted sum of the following loss.

$$L_{total} = (E_{id}, E^s nid, E^t_{nid}, D_{dom}, D_{img}, G^s, G^T) = \lambda_{cyc} * L_{cyc} + \lambda^s_{id} * L^s_{id} + \lambda^t_{id} * L^t_{id} + \lambda_{dom} * L_{dom} + \lambda_{img} * L_{img} \tag{4.13}$$

We have followed common way to give more weight to $\lambda_{cyc} = 2$ to follow image to image translation techniques [11].We generate pseudo labels for target domain every two epochs. We also followed alternative updating policy in training GAN to train first $E_i d, E^s_{nid}, E^t_{nid}, G^s, G^t$ and then train D_{dom}, D_{img}.

4.6 Implementation Details

To implemented entire architecture we have used pytorch Deep learning framework. For E_{id} we have used resnet-50 based architecture which was pre-trained on ImageNet. For E^s_{nid} and E^t_{id} share same architecture with four convolution layers followed four residuals layers. For G^s and

Algorithm 1 Training VehicleNet Architecture using Meta Learning

$D_1 \leftarrow AICITY(Labelled)$.$\mathrm{D}_2 \leftarrow VRIC(Labelled)$.
$D_3 \leftarrow Veri-776(Unlabelled)$. Here the dataset$\mathrm{D}_1$ and D_2 is Meta-Train and D_3 is Meta-Test.
Intialize θ parameter for $F_\theta(E_{id}, E^S_{nid}, E^T_{nid}, G^S, G^T, D_{img}, D_{dom})$.
β :Innerloop learning rate.
α :Outerloop learning rate.
n_{epochs} : no of epochs.
max_{ietr} : No of iteration.
procedure VehicleReidMetaLearning(D_1, D_2, D_3))
  for `iter in` max_{iter} **do**
    Apply pseudo labeling on D_3
    for `epoch in` n_{epochs} **do**
      for D^i `in` `[`D^1, D^2`]` **do**
        sample k data points from D^i and D^3 for Meta Train
        $\nabla\hat{\theta_i}L \leftarrow \mathrm{F}^{VehicleReid}_{\hat{\theta}}(E_{app}, E^S_{str}, E^T_{str}, G^S, G^T, D_{img}, D_{dom}, D^i_k, D^3_k)$
        $\hat{\theta}_i \leftarrow$ - $\beta\nabla\hat{\theta}_i L$
        sample M data points from D_i and D_3 for meta update
      end for
      $\theta \leftarrow \theta - \alpha \sum_{i=1}^{2} \nabla F^{VehicleReid}_{\hat{\theta}_i}(E_{app}, E^S_{str}, E^T_{str}, G^S, G^T, D_{img}, D_{dom}, D^i_M, D^3_M)$
    end for
  end for
  Return θ

G^t follows same architecture which has four residual block [7] followed by four convolution layers. Each residual block consist of two adaptive instance normalization layers [10].D_{img} used popular patchGAN [22] and for D_{dom} is used four fully connected multi-layer perceptron. Comming to different gradient optimizer we have used Adam optimizer for all of our components. As we are using DBSCAN to generate pseudo labels for target images so for that we have taken $eps = 0.45$ and for $minpts = 7$.

Chapter 5

Result and Analysis

5.1 Performance Metric

To evaluate the efficacy of our architecture we have used two different performance metric one is mAP(mean Average precision) and another one is CMC(Cumulative matching characteristics) in other words also called Rank. In mAP first we calculate the average precision for each query and then we take average of all the query.

$$\begin{aligned} Avg(q) &= \frac{1}{GTP}\sum_{k}^{n} p@k * rel@k \\ mAP &= \frac{\sum_{q}^{Q} avg(q)}{Q} \end{aligned} \tag{5.1}$$

Now to calculate CMC, given an query image, algorithm will rank all the gallery image according to their distance to the query image from small to large and the CMC top-k accuracy is calculated as

$$Acc_k = \begin{cases} 1, & \text{if top-k ranked gallery samples contain the query idenity .} \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

5.2 Quantitative Analysis

In the quantitative analysis we have performed both comparison based analysis with different state of the art and also perform the ablation studies to check the contribution of all different components of the architecture.

5.2.1 Ablation Study

In ablation studies we have four different variants, variant one we obtain by train E_{id} on the source domain only test on the target domain. As expected the variant one has the lowest

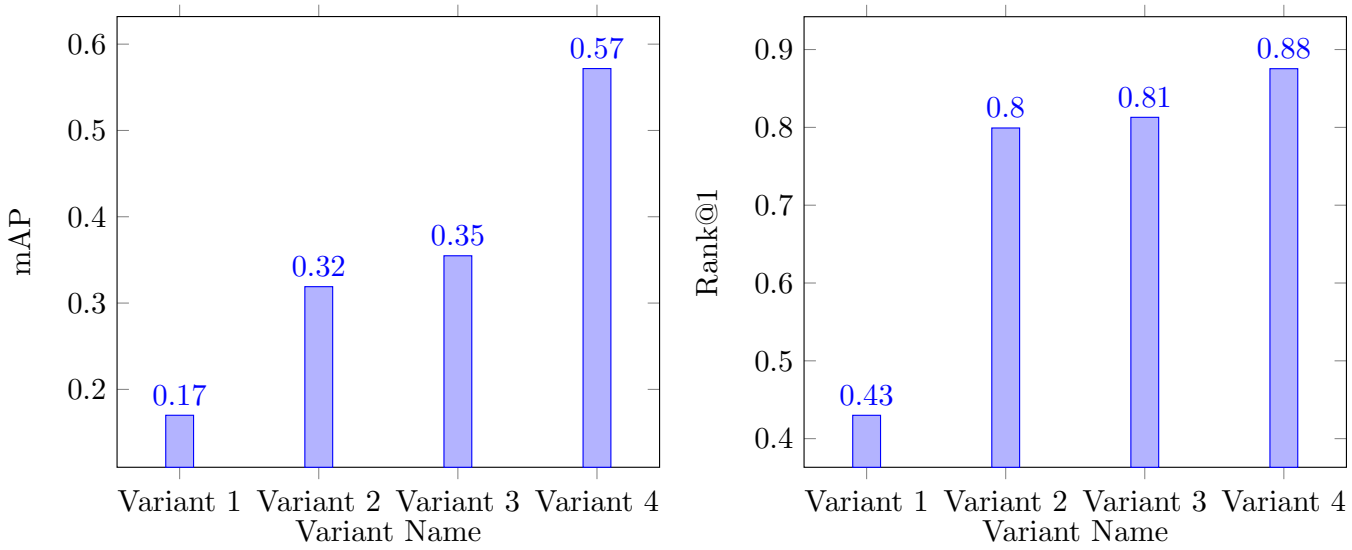
performance compared to all other variant.In variant two, we have integrated pseudo labeling and domain adversarial models as expected, the using the domain adversarial and pseudo labeling modules which force the adaption module to adapt on the shared id-space.In variant three, we have integrated Meta learning algorithm to our final architecture and its giving an edge to the performance of our architecture . In the final variant which Variant four, we have fine tuned one of the pre-trained model which is based on vehicle re-identification and we achieved highest performance compared to all other variants.

Variant Name	Baseline	Pseudo Labels	Meta Learning	CTACL Finetune
Variant 1	✓	✗	✗	✗
Variant 2	✓	✓	✗	✗
Variant 3	✓	✓	✓	✗
Variant 4	✓	✓	✓	✓

Table 5.1: Different variations of ablation studies

Variant Name	mAP	Rank@1	Rank@5	Rank@10
Variant 1	0.17	0.43	0.60	0.79
Variant 2	0.319	0.7992	0.87	0.905
Variant 3	0.3548	0.8129	0.9011	0.9237
Variant 4	0.5717	0.8754	0.9333	0.9476

Table 5.2: Results of different ablation studies

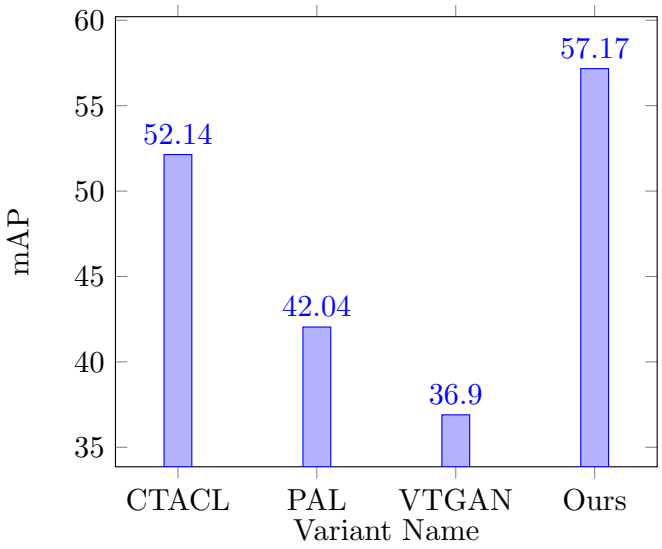


5.2.2 Comparioson with state-of-the-art

We exclusively evaluated our architecure with different state-of-the-art methods. We also have explained the methods that they are follow at Related work section. We have done experiment only on the Veri-776 dataset only. Follow are the result that we achieved

Architecture	mAP	Rank@1	Rank@5	Rank@10
CTACL [30]	52.14	88.38	93.27	95.11
PAL [23]	42.05	68.17	79.91	-
VTGAN [23]	36.9	77.8	88.5	-
Vehicle Meta Learning(ours)	57.17	87.54	93.33	94.76

Table 5.3: Comparison with different state-of-art



5.3 Qualitative Analysis

In qualitative analysis, we wanted to show the retrieval results given the query images at Figure 5.1 and how iteration wise the synthetic images are getting generating at Figure 5.2

Figure 5.1: Given a query image from veri-776 dataset , our model is extracting all the gallery image correctly

Figure 5.2: Given target image and source image how our model generating the synthetic images, for example here the generated synthetic image have target id-features and source non-id features which is here color

Figure 5.3: Result of qualitative analysis

Chapter 6

Conclusion

Here we proposed joint learning framework which disentangle id-related features and id-unrelated features and adaptation happened exclusively on the id-related features space.Domain adaptation followed by pseudo labeling and cross domain cyclic image generation promote each other during joint learning.We perform experiments on VeRI-776, and compare against various unsupervised techniques to show the efficacy of our model.

Bibliography

[1] Chang, Y.-L., Liu, Z. Y., Lee, K.-Y., and Hsu, W. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9066–9075.

[2] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems 29* (2016).

[3] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.

[4] Fan, H., Zheng, L., Yan, C., and Yang, Y. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14*, 4 (2018), 1–18.

[5] Finn, C., Xu, K., and Levine, S. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems 31* (2018).

[6] Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., and Huang, T. S. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 6112–6121.

[7] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[8] He, S., Luo, H., Chen, W., Zhang, M., Zhang, Y., Wang, F., Li, H., and Jiang, W. Multi-domain learning and identity mining for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 582–583.

[9] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework.

[10] Huang, X., and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 1501–1510.

[11] Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 172–189.

[12] Huang, Y., Liang, B., Xie, W., Liao, Y., Kuang, Z., Zhuang, Y., and Ding, X. Dual domain multi-task model for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems* (2020).

[13] Kanaci, A., Zhu, X., and Gong, S. Vehicle re-identification in context. In *German Conference on Pattern Recognition* (2018), Springer, pp. 377–390.

[14] Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 35–51.

[15] Li, M., Huang, X., and Zhang, Z. Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 194–204.

[16] Lin, Y., Xie, L., Wu, Y., Yan, C., and Tian, Q. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3390–3399.

[17] Liu, C.-T., Lee, M.-Y., Wu, C.-W., Chen, B.-Y., Chen, T.-S., Hsu, Y.-T., Chien, S.-Y., and Center, N. I. Supervised joint domain learning for vehicle re-identification. In *CVPR Workshops* (2019), pp. 45–52.

[18] Liu, X., Liu, W., Ma, H., and Fu, H. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE international conference on multimedia and expo (ICME)* (2016), IEEE, pp. 1–6.

[19] Liu, X., Liu, W., Mei, T., and Ma, H. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision* (2016), Springer, pp. 869–884.

[20] Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning* (2015), PMLR, pp. 97–105.

[21] Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. *Advances in neural information processing systems 29* (2016).

[22] Naphade, M., Wang, S., Anastasiu, D. C., Tang, Z., Chang, M.-C., Yang, X., Yao, Y., Zheng, L., Chakraborty, P., Lopez, C. E., et al. The 5th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4263–4273.

[23] Peng, J., Wang, Y., Wang, H., Zhang, Z., Fu, X., and Wang, M. Unsupervised vehicle re-identification with progressive adaptation. *arXiv preprint arXiv:2006.11486* (2020).

[24] Shen, Y., Xiao, T., Li, H., Yi, S., and Wang, X. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1900–1909.

[25] Sun, B., and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision* (2016), Springer, pp. 443–450.

[26] Wang, D., and Zhang, S. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 10981–10990.

[27] Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., Yan, J., Wang, S., Li, H., and Wang, X. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 379–387.

[28] Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., Yan, J., Wang, S., Li, H., and Wang, X. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 379–387.

[29] Wei, L., Zhang, S., Gao, W., and Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 79–88.

[30] Yu, J., Kim, J., Kim, M., and Oh, H. Camera-tracklet-aware contrastive learning for unsupervised vehicle re-identification. *arXiv preprint arXiv:2109.06401* (2021).

[31] Yu, J., and Oh, H. Unsupervised vehicle re-identification via self-supervised metric learning using feature dictionary. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2021), IEEE, pp. 3806–3813.

[32] Zheng, Z., Jiang, M., Wang, Z., Wang, J., Bai, Z., Zhang, X., Yu, X., Tan, X., Yang, Y., Wen, S., et al. Going beyond real data: A robust visual representation for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 598–599.