Unsupervised Domain Adaptation for Real World Person Re-Identification

Tiago de Carvalho Gallo Pereira

Departamento de Ciência da Computação, Universidade de Brasília, Brasília - DF

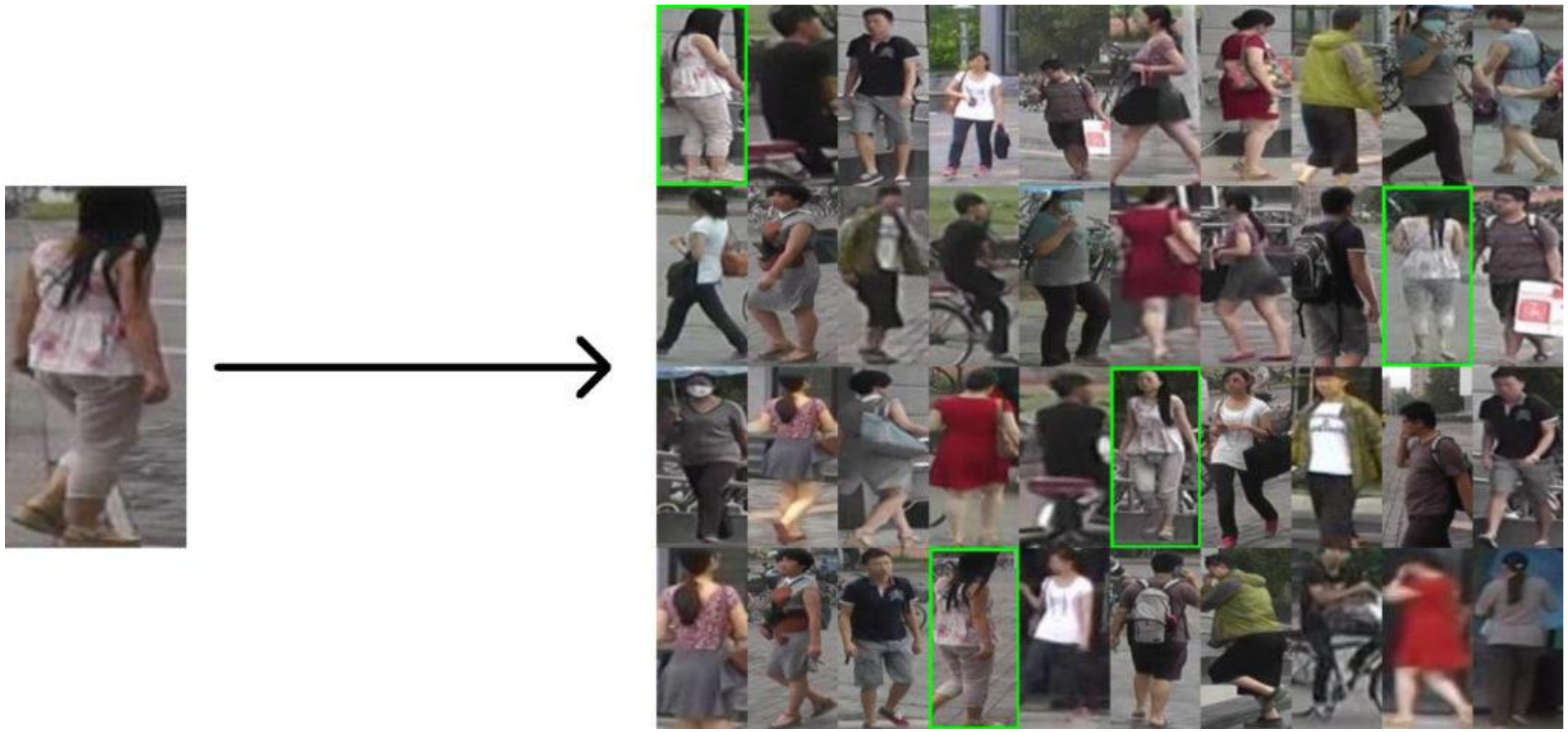
Examination Board:

- Teófilo E. de Campos (Supervisor – UnB)
- Krystian Mikolajczyk (Imperial College London)
- Bruno L. Macchiavello Espinoza (UnB)
- Flávio de Barros Vidal (substitute - UnB)

April 2022

Definition

Person Re-Identification is an image retrieval task, where the object in the images are people.



Motivation

The goal of person Re-ID is **Matching person images from different non-overlapping cameras views**. However, the addition of a new viewpoint usually impact the algorithm performance.

Objectives

In this work, we aim to create person Re-ID framework capable of learning robust representations from non-annotated data. To achieve that, we set 3 auxiliary goals:

1. Implement a baseline domain adaptation method to start from;
2. Identify the flaws in our baseline domain adaptation method and propose techniques to undermine them;
3. Improve our proposed methods and compare them with the state-of-the-art algorithms.

Contributions

While working towards our goals, we proposed some techniques that generated the following publications:

- Pereira, T. and de Campos, T. Domain Adaptation for Person Re-identification on New Unlabeled Data (**best student paper award winner** at VISAPP 2020) [1]

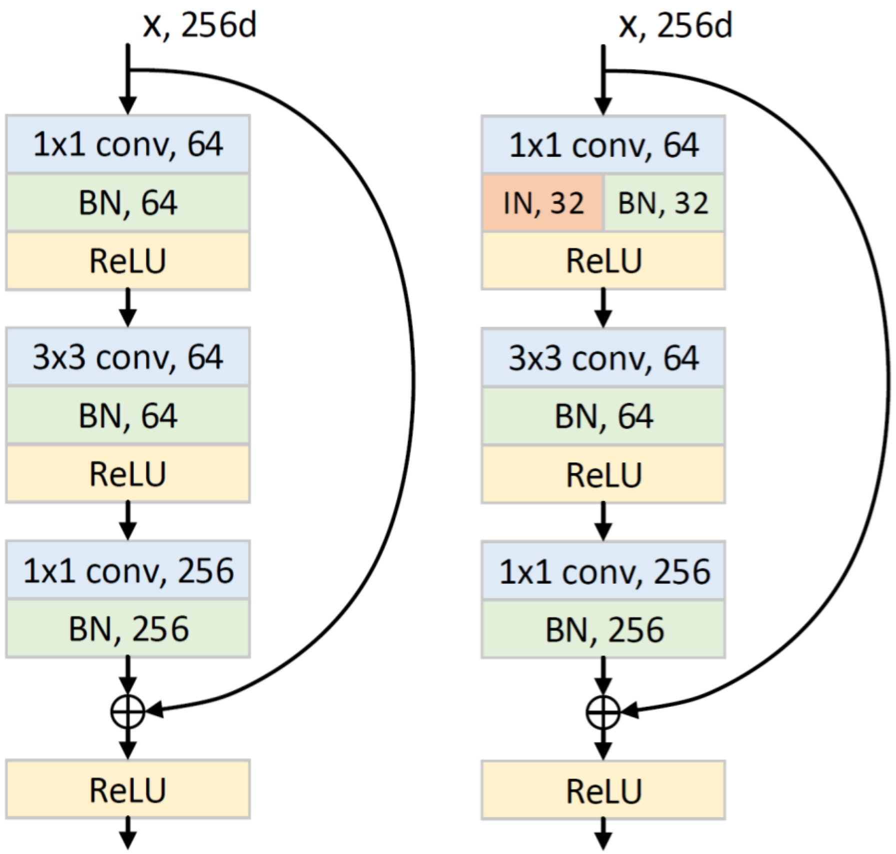
- Pereira, T. and de Campos, T. Domain adaptation for person re-identification on new unlabeled data using AlignedReID++ (IJPRAI) [2]

- Pereira, T. and de Campos, T. Learn by Guessing: Multi-Step Pseudo-Label Refinement for Person Re-Identification (VISAPP 2022) [3]

Model Architectures – Resnet 50 [4] and IBN Net-50 a [5]

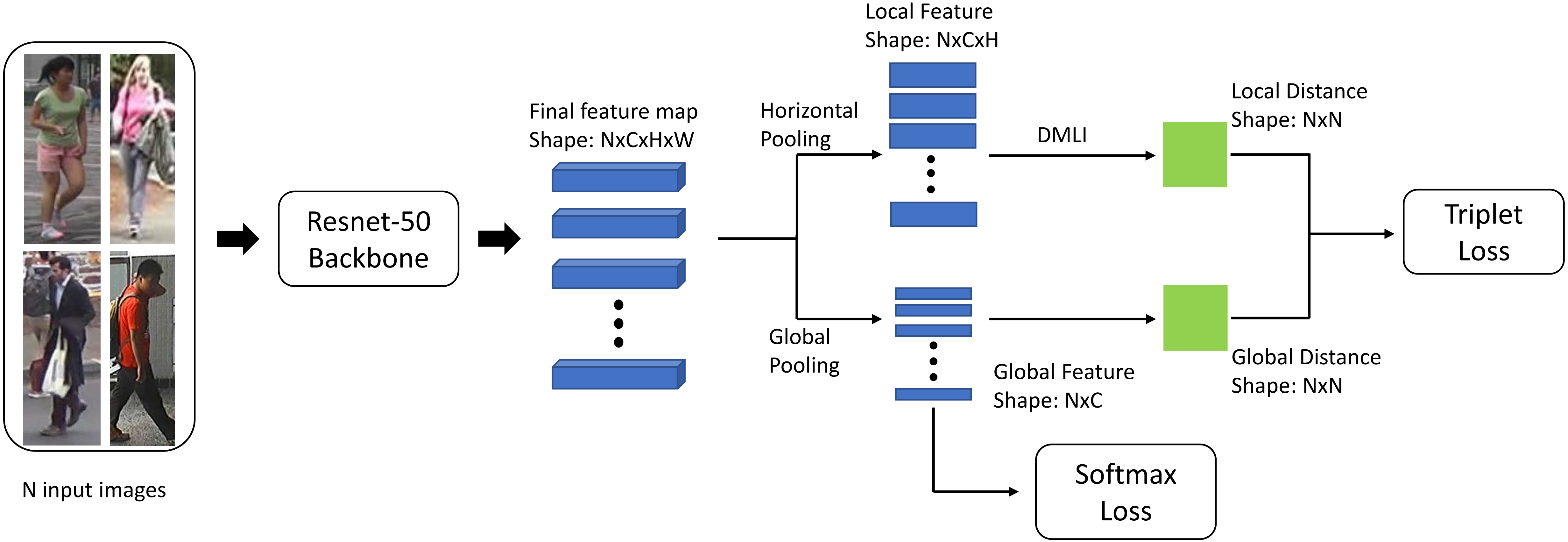
For the person Re-ID challenge, we need a model architecture that can:

- Encode the person information into a feature vector
- Take advantage from information of multiple semantic levels
- Disregard background information
- Be robust against variations in illumination, angle, saturation, resolution, distance from people.



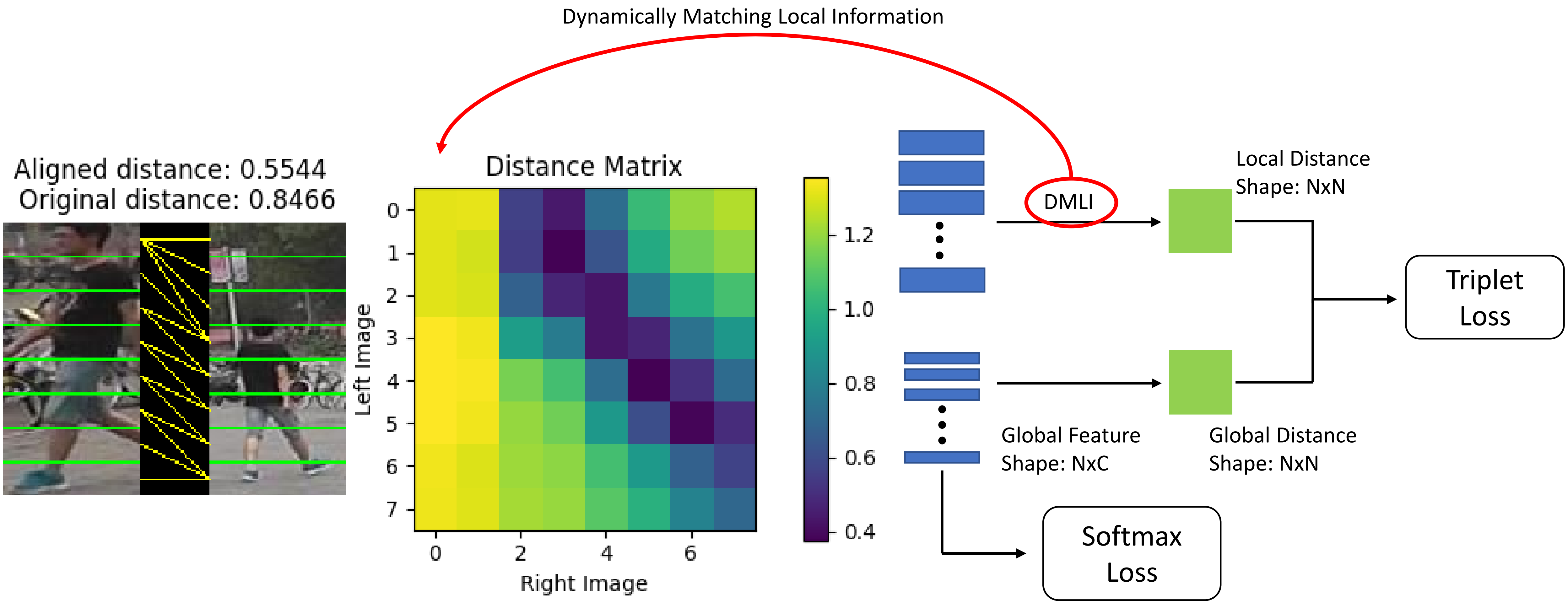
[4] He, K. et al.: Deep Residual Learning for Image Recognition. CVPR, 2016.
[5] Pan, X. et al.: Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. ECCV, 2018.

Model Architectures – Aligned ReID++ [6]

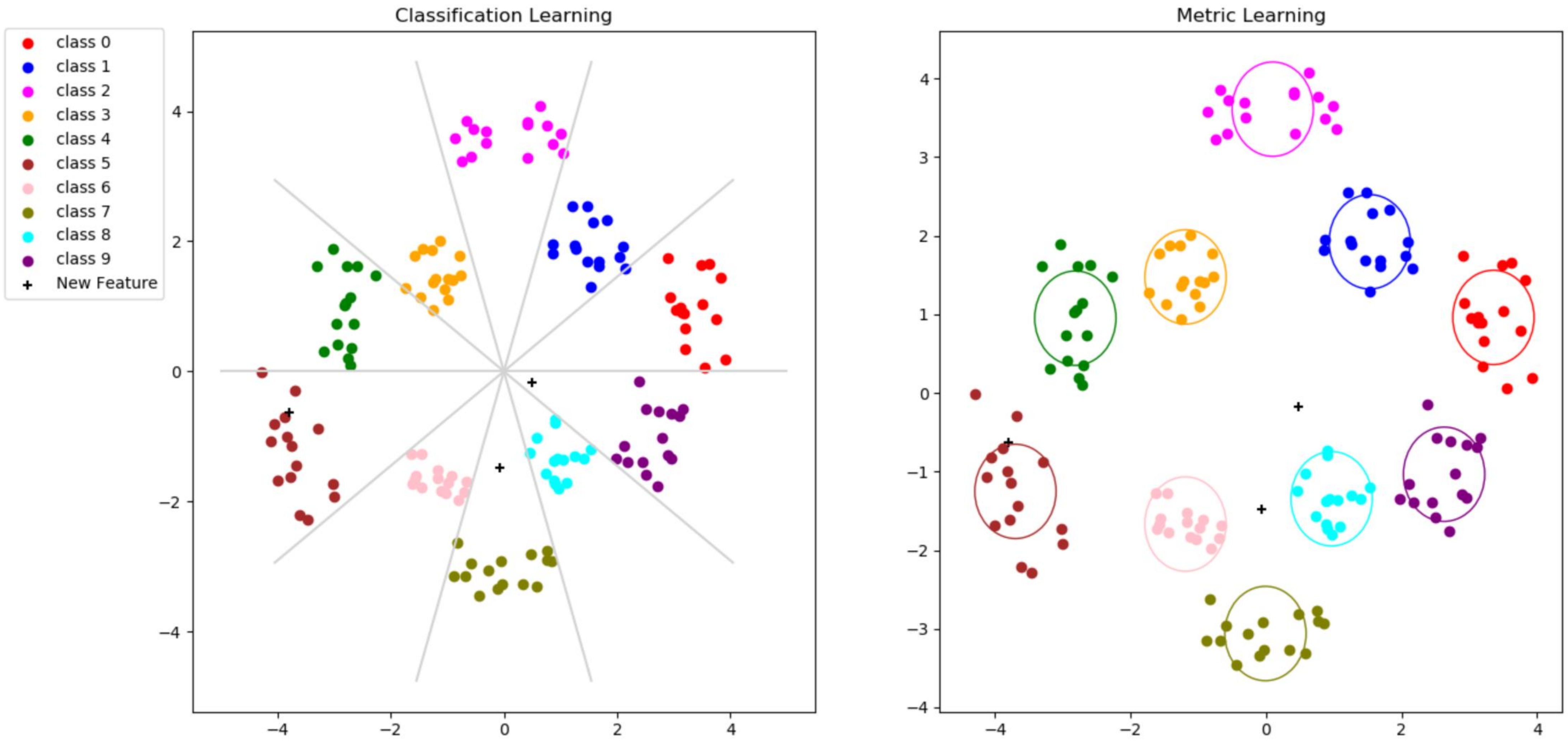


[6] Luo, H. et al. : AlignedReID++: Dynamically matching local information for person re-identification. Pattern Recognition, 2019.

Model Architectures – Aligned ReID++

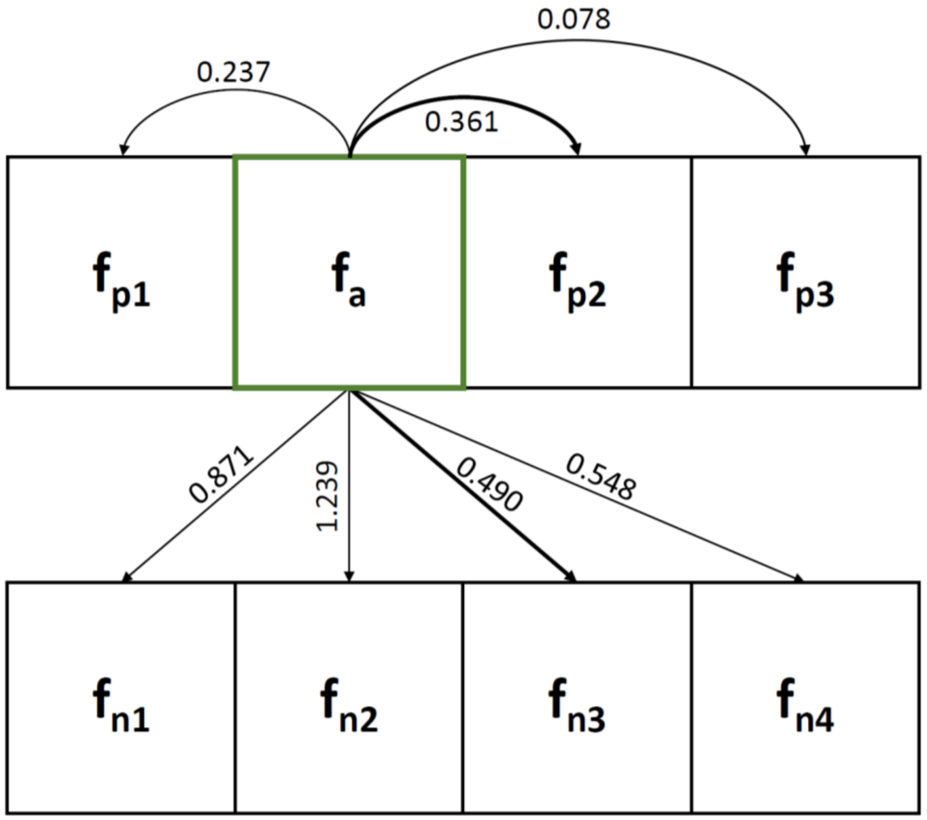


Metric Learning vs. Classification Learning



Triplet Loss & Batch hard

- The Triplet Loss is responsible for producing output vectors that belong to a Euclidean feature space;
- It is better than the contrastive loss, once it can push pairs from different people away while pulling feature pairs from same people together;
- **Challenge:** How to choose the best triplets? Based on Hermans et al. [7] work batch hard is the best approach.



[7] Hermans, A., Beyer, L., and Leibe, B. In defense of the triplet loss for person re-identification. arXiv 2017.

Viper Dataset [8]

	Viper
Release Year	2007
Samples	1264
Identities	632
Cameras	2
Avg Number of Cameras Passed per Identity	2
Scene	outdoor

[8] Gray, D. et al.: Evaluating appearance models for recognition, reacquisition, and tracking. In In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, 2007.

CUHK03 Dataset [9]

	CUHK03
Release Year	2014
Samples	28192
Identities	1467
Cameras	2
Avg Number of Cameras Passed per Identity	2
Scene	indoor

[9] Li, W. et al.: DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. CVPR,2014.

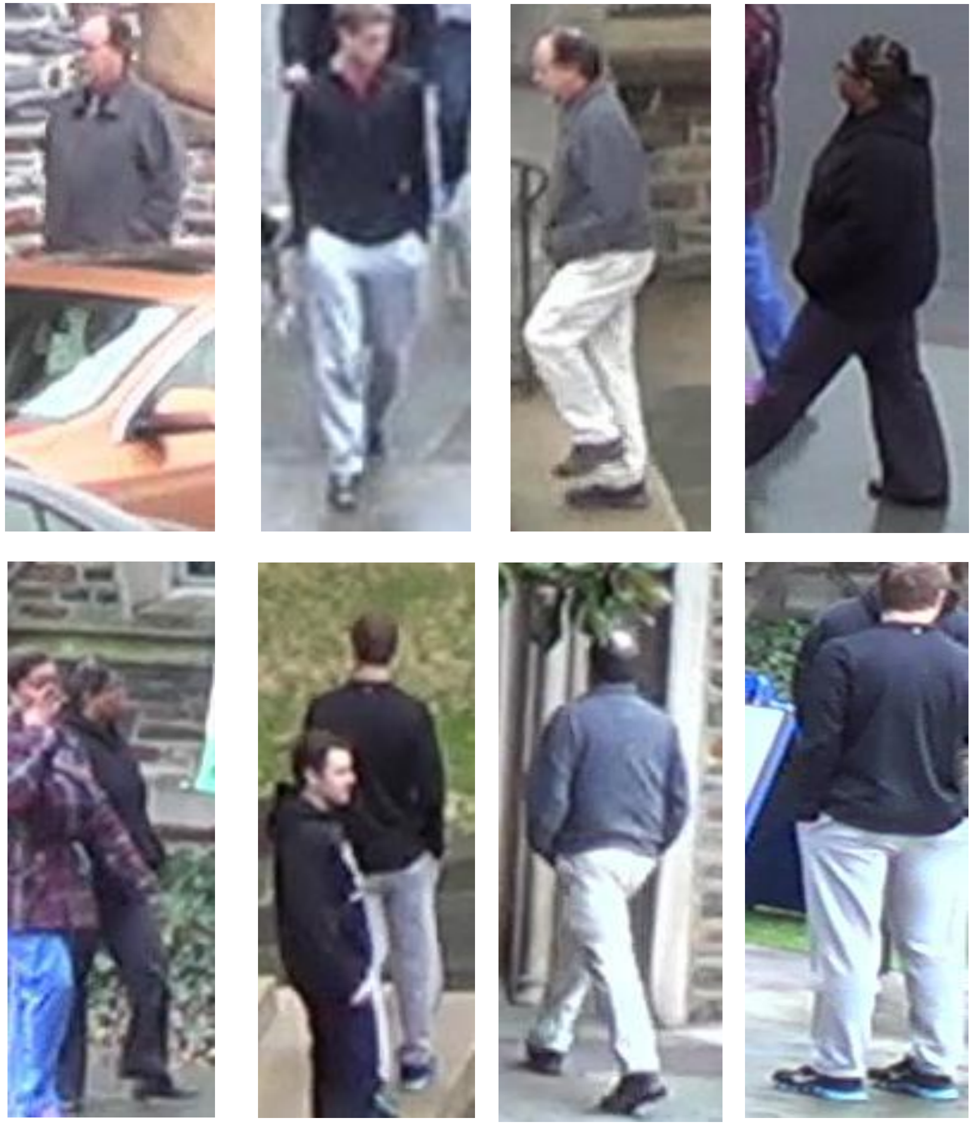
Market1501 Dataset [10]

	Market1501
Release Year	2015
Samples	32668
Identities	1501
Cameras	6
Avg Number of Cameras Passed per Identity	4.42
Scene	outdoor

[10] Zheng, et al.: Scalable Person Re-identification: A Benchmark. ICCV, 2015.

DukeMTMC Dataset [11]

	DukeMTMC
Release Year	2016
Samples	36411
Identities	1812
Cameras	8
Avg Number of Cameras Passed per Identity	2.67
Scene	outdoor



[11] Zheng, Z. et al.: Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. ICCV, 2017.

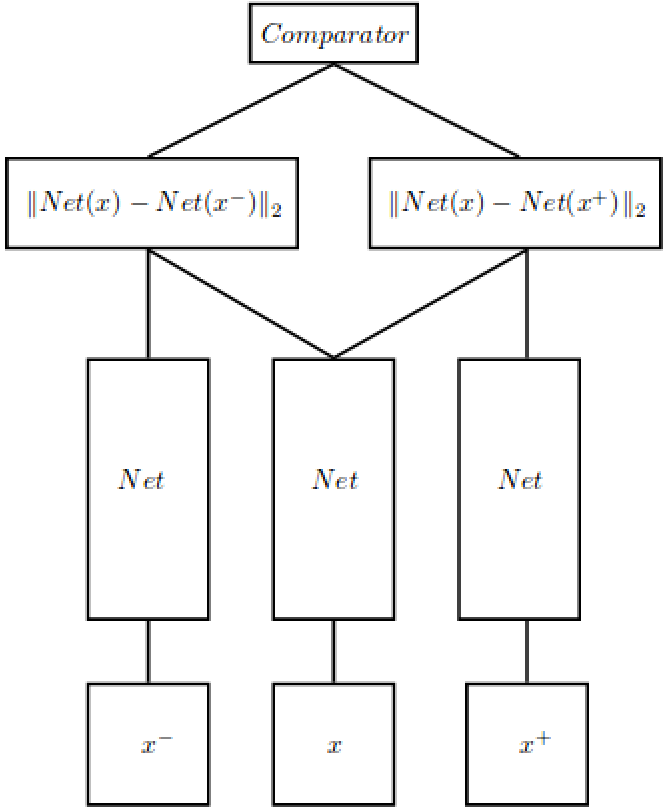
Overview

	Viper	CUHK03	Market1501	DukeMTMC
Release Year	2007	2014	2015	2016
Samples	1264	28192	32668	36411
Identities	632	1467	1501	1812
Cameras	2	2	6	8
Avg Number of Cameras Passed per Identity	2	2	4.42	2.67
Scene	outdoor	indoor	outdoor	outdoor

Training Strategy

Our general training strategy had the following configurations:

- ResNet-50 or AlignedReID++
- Triplet loss with batch hard
- Adam optimizer
- Batch scheduler

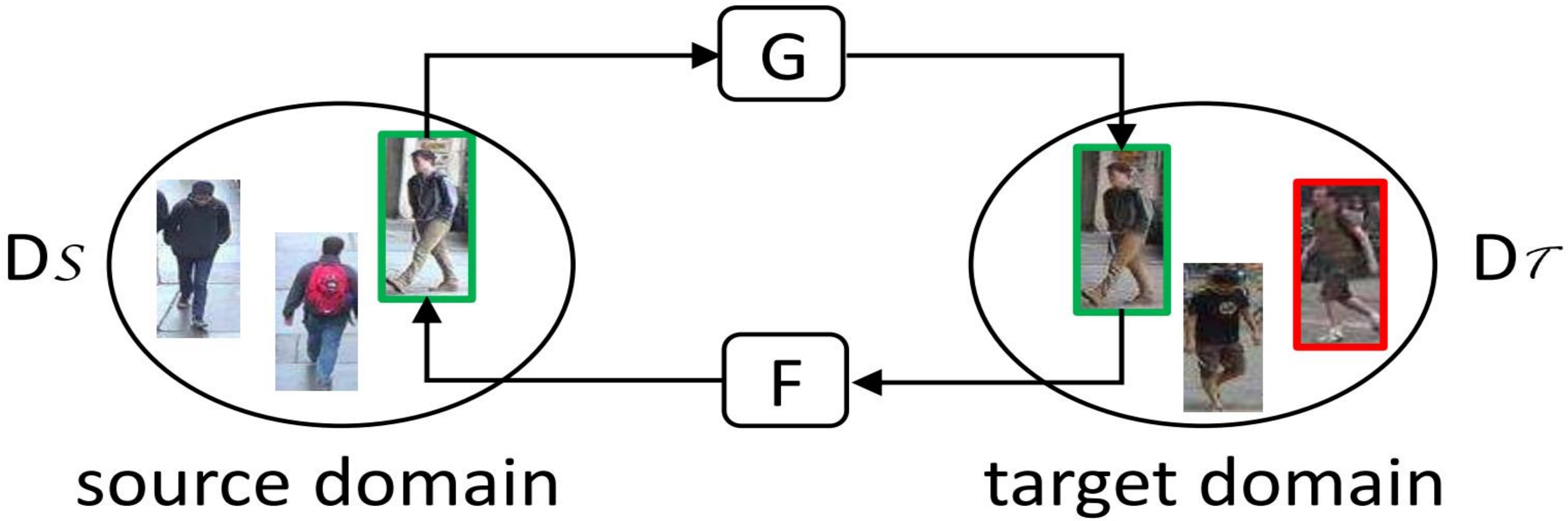


Algorithm 1 Batch Scheduler

$\gamma = 2 \times \tau$
for $i = 0$ to *epochs* **do**
    $loss = train(i, \gamma)$
    if $loss < (0.8 \times m)$ **then**
        $\gamma = \gamma \times 2$
    end if
end for

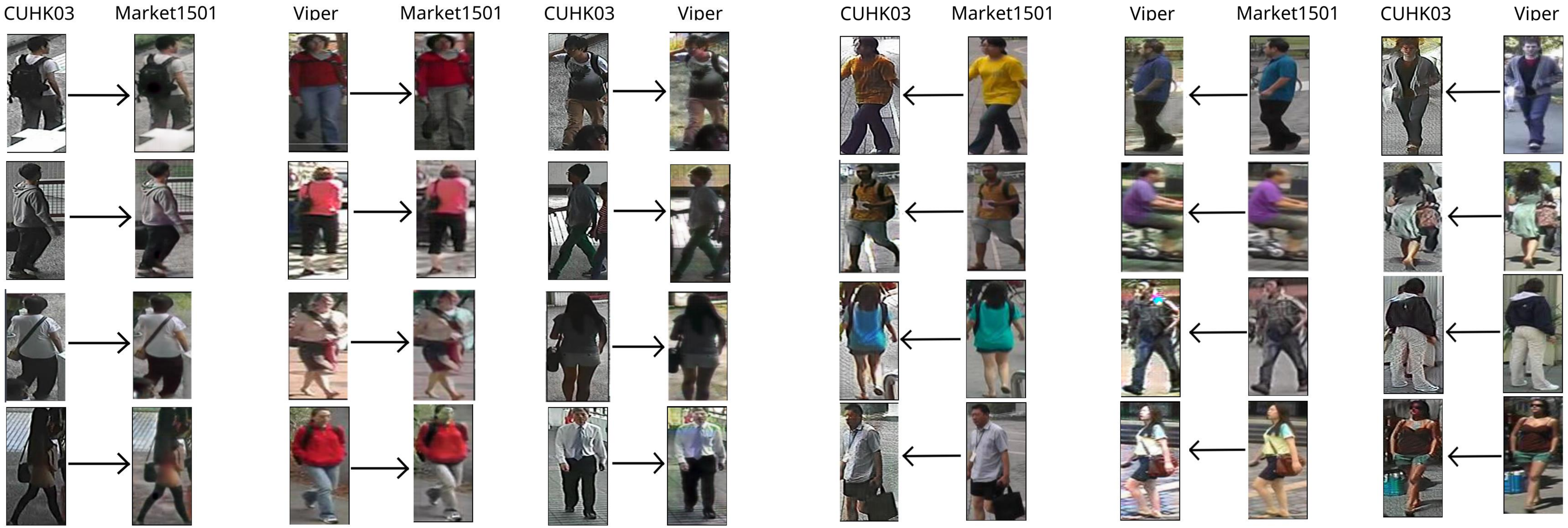
Domain Adaptation – Cycle GAN Step

Train CycleGAN [12] with images from source and target domain → Generate Intermediate Dataset → Fine Tune the CNN using the intermediate dataset



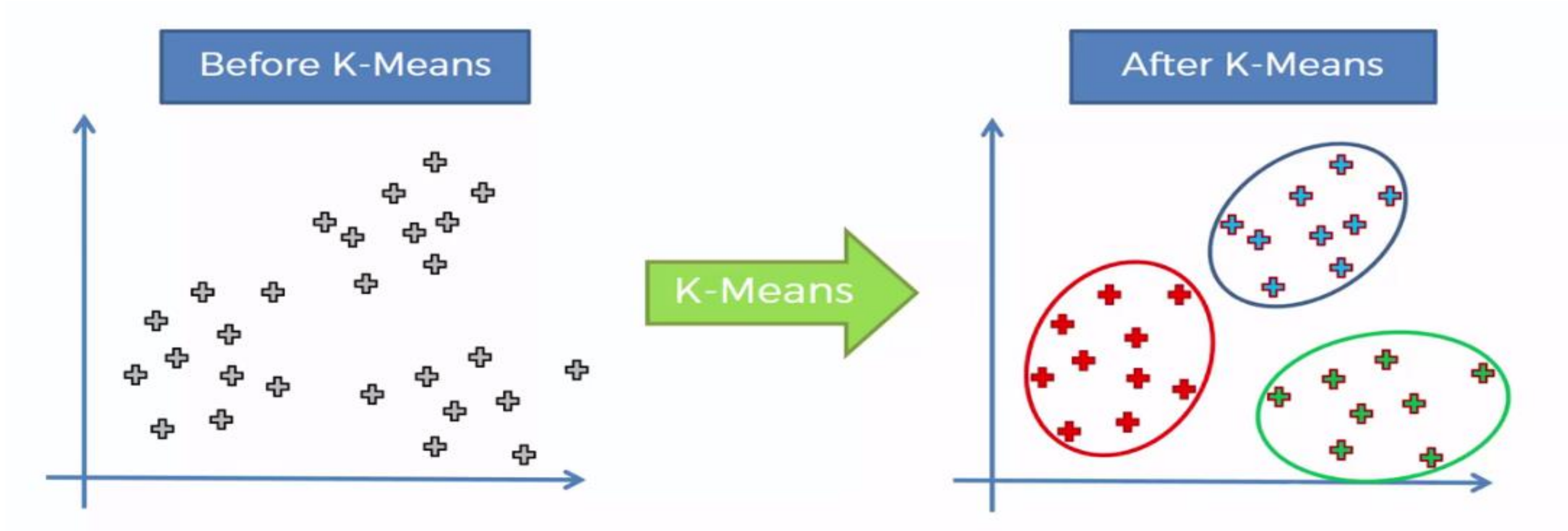
[12] Isola, P. et al.: Image-To-Image Translation With Conditional Adversarial Networks. CVPR, 2017.

Domain Adaptation – Cycle GAN Results



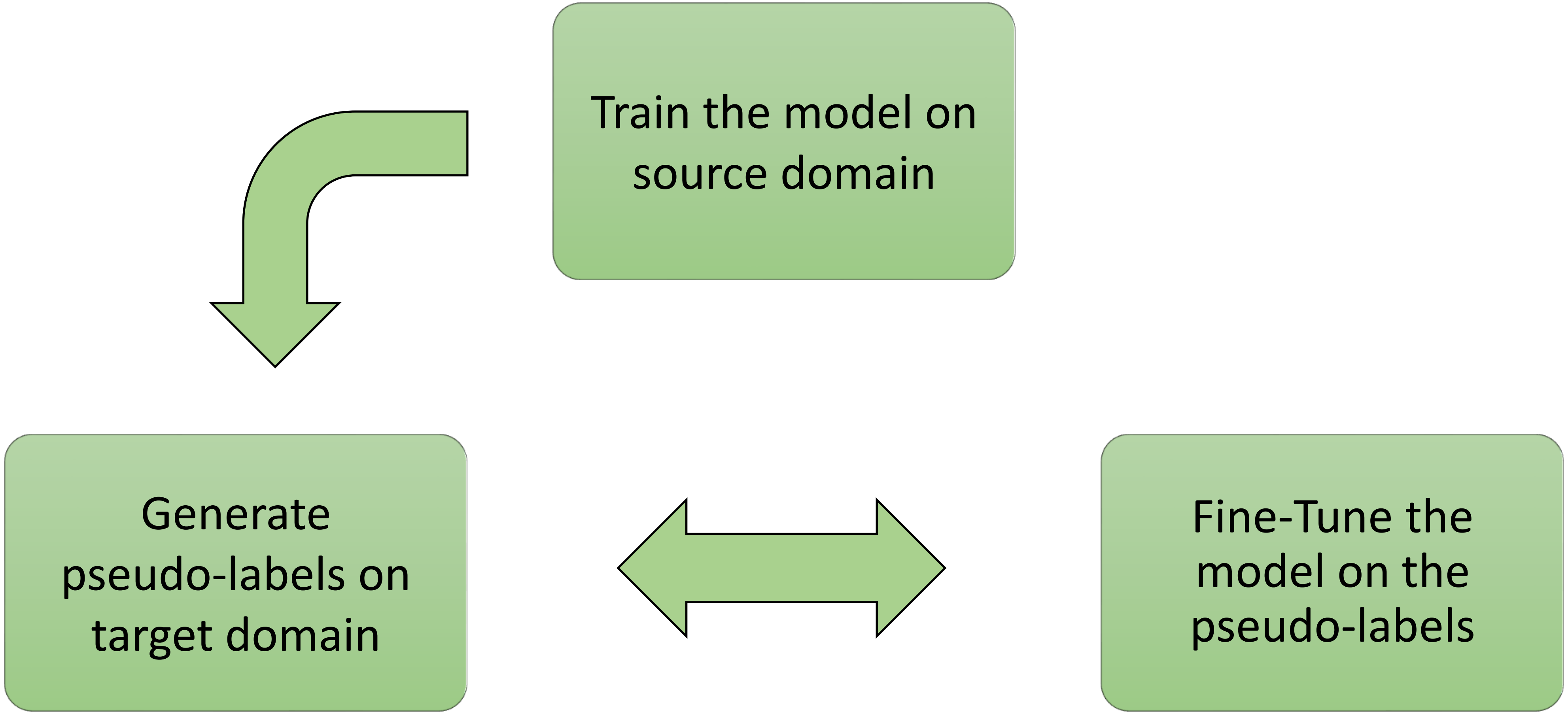
Domain Adaptation – Pseudo-Labels Step

Extract Features from target domain images → Group these features using k-means [13] → Generate a pseudo labelled dataset with the grouped features → Fine Tune the CNN using the pseudo labelled dataset



[13] Hartigan, J. A. and Wong, M. A. A K-means clustering algorithm. In: Journal of the Royal Statistical Society 1979.

Progressive Learning [14]



[14] Fan, H. et al.: Unsupervised Person Re-identification: Clustering and Fine-tuning. TOMM, 2018.

Domain Adaptation – Pseudo-Labels Results

Without Progressive learning

With Progressive learning

Results - Baseline

			Accuracy (CMC scores)		
Source Domain	**Target Domain**	**Method**	**Rank – 1**	**Rank - 5**	**Rank - 10**
Market 1501	**Viper**	Direct Transfer	12.5%	25.0%	33.1%
		CycleGAN	9.8%	26.9%	36.4%
		Ours	**13.9%**	**29.0%**	**40.7%**
	CUHK 03	Direct Transfer	19.9%	49.4%	63.2%
		CycleGAN	34.8%	66.7%	79.1%
		Ours	**38.2%**	**69.7%**	**81.6%**
CUHK 03	**Viper**	Direct Transfer	10.1%	22.5%	29.0%
		CycleGAN	11.6%	25.5%	34.7%
		Ours	**13.6%**	**33.9%**	**46.0%**
	Market 1501	Direct Transfer	26.8%	45.9%	55.1%
		CycleGAN	35.8%	56.5%	65.7%
		Ours	**37.3%**	**60.4%**	**70.4%**
Viper	**CUHK 03**	Direct Transfer	5.9%	18.1%	29.0%
		CycleGAN	31.9%	64.4%	77.5%
		Ours	**36.1%**	**69.2%**	**81.3%**
	Market 1501	Direct Transfer	5.7%	15.5%	22.2%
		CycleGAN	6.7%	17.0%	23.7%
		Ours	**8.6%**	**20.5%**	**28.4%**

Results – AlignedReID++

			Accuracy (CMC scores)		
Source Domain	**Target Domain**	**Method**	**Rank – 1**	**Rank – 5**	**Rank - 10**
Market 1501	**Viper**	Direct Transfer	22.9%	**41.8%**	50.0%
		CycleGAN	21.4%	40.2%	50.3%
		Ours	**23.7%**	41.5%	**50.8%**
	CUHK 03	Direct Transfer	22.5%	45.0%	47.2%
		CycleGAN	37.0%	69.1%	80.9%
		Ours	**42.9%**	**72.5%**	**81.2%**
CUHK 03	**Viper**	Direct Transfer	20.6%	38.0%	47.2%
		CycleGAN	21.8%	**43.2%**	52.2%
		Ours	**22.5%**	**43.2%**	**54.1%**
	Market 1501	Direct Transfer	38.7%	55.1%	62.6%
		CycleGAN	42.7%	59.7%	67.3%
		Ours	**46.8%**	**65.9%**	**73.6%**
Viper	**CUHK 03**	Direct Transfer	9.9%	27.9%	40.1%
		CycleGAN	17.1%	41.6%	55.8%
		Ours	**20.4%**	**43.9%**	**58.5%**
	Market 1501	Direct Transfer	15.9%	28.2%	35.4%
		CycleGAN	23.1%	37.9%	45.8%
		Ours	**28.4%**	**46.4%**	**55.2%**

Results – Ablation Studies – Batch Scheduler

				Accuracy (CMC scores)		
Source Domain	**Target Domain**	**Method**	**Batch Scheduler**	**Rank – 1**	**Rank – 5**	**Rank - 10**
Market1501	**CUHK03**	CycleGAN	✗	37.0%	69.1%	80.9%
			✓	**38.9%**	**69.2%**	**81.1%**
		Ours	✗	42.9%	72.5%	81.2%
			✓	**43.1%**	**72.7%**	**84.2%**
CUHK03	**Market1501**	CycleGAN	✗	**42.7%**	**59.7%**	**67.3%**
			✓	38.4%	57.2%	65.5%
		Ours	✗	46.8%	65.9%	73.6%
			✓	**50.1%**	**68.2%**	**75.6%**

Results – Ablation Studies – CycleGAN

			Accuracy (CMC scores)		
Source Domain	**Target Domain**	**CycleGAN**	**Rank – 1**	**Rank – 5**	**Rank - 10**
Market1501	**Viper**	✗	21.5%	38.3%	80.9%
		✓	**23.7%**	**41.5%**	**81.1%**
	CUHK03	✗	31.6%	58.5%	81.2%
		✓	**43.1%**	**72.7%**	**84.2%**
CUHK03	**Viper**	✗	19.5%	41.0%	67.3%
		✓	**22.5%**	**43.2%**	**65.5%**
	Market1501	✗	45.7%	61.5%	73.6%
		✓	**50.1%**	**68.2%**	**75.6%**
Viper	**CUHK03**	✗	18.0%	40.8%	53.6%
		✓	**20.4%**	**43.9%**	**58.5%**
	Market1501	✗	23.0%	37.6%	44.9%
		✓	**28.4%**	**46.4%**	**55.2%**

Results – Ablation Studies – Progressive Learning

			Accuracy (CMC scores)		
Source Domain	**Target Domain**	**PL Iterations**	**Rank – 1**	**Rank – 5**	**Rank - 10**
Market1501	**Viper**	1	**23.7%**	**41.5%**	**50.8%**
		2	18.2%	36.9%	46.0%
	CUHK03	1	43.1%	72.7%	84.2%
		3	**47.8%**	**75.9%**	**84.2%**
CUHK03	**Viper**	1	**22.5%**	**43.2%**	**54.1%**
		2	20.7%	40.8%	50.6%
	Market1501	1	50.1%	68.2%	75.6%
		9	**64.3%**	**81.5%**	**87.5%**
Viper	**CUHK03**	1	20.4%	43.9%	58.5%
		14	**51.2%**	**76.2%**	**83.8%**
	Market1501	1	28.4%	46.4%	55.2%
		14	**55.2%**	**73.9%**	**81.0%**

Multi-Step Pseudo-Label Refinement

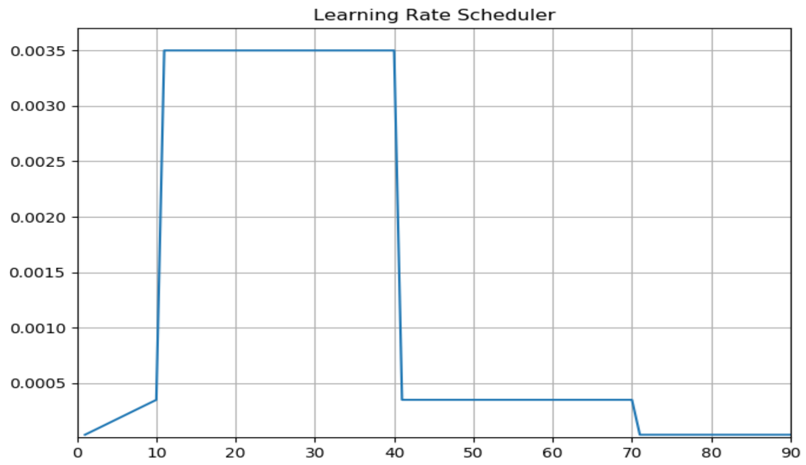
We identified some problems the previous method that we aim to solve using our Multi-Step Pseudo-Label Refinement method.

- The model lack of generalization;
- The noisy and low-quality pseudo-labels;
- The high influence of camera characteristics in the pseudo-labels generation;
- The high computational cost to train GANs and generate the intermediate dataset.

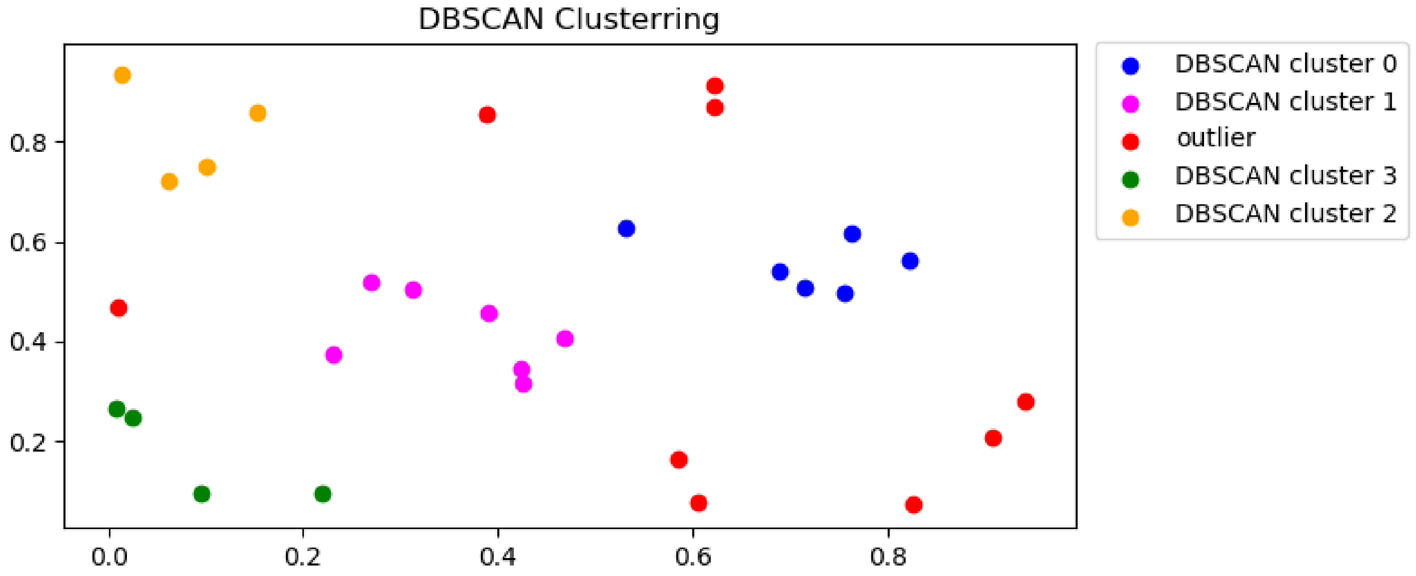
Step 1 – Architecture + Training Strategy

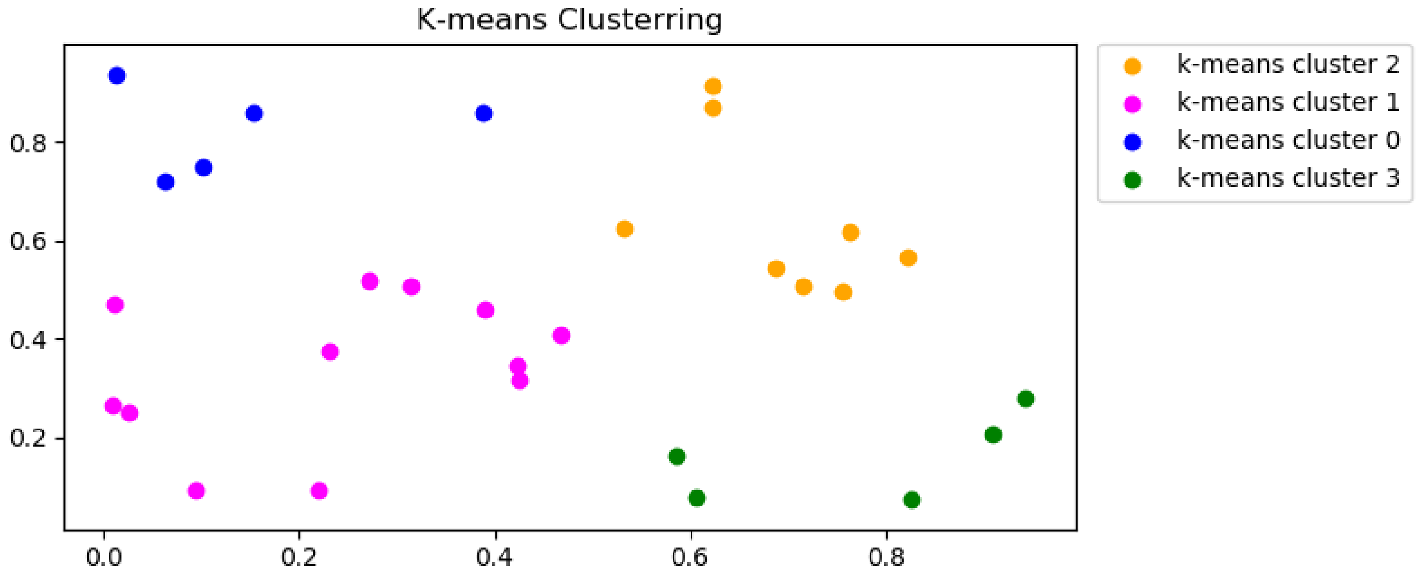
Our general training strategy had the following configurations:

- IBN Net-50 a
- Adam optimizer
- A three-factor loss function given by $\mathcal{L} = \mathcal{L}_{triplet} + \mathcal{L}_{ID} + 0.005 * \mathcal{L}_{center}$ where:
  - $\mathcal{L}_{triplet}$ is the triplet Loss responsible for the metric leaning,
  - $\mathcal{L}_{ID}$ is a label smooth cross entropy loss for person ID classification
  - $\mathcal{L}_{center}$ is a center loss to guarantee cluster compactness
- A learning rate scheduler for the 90 training epochs defined by:

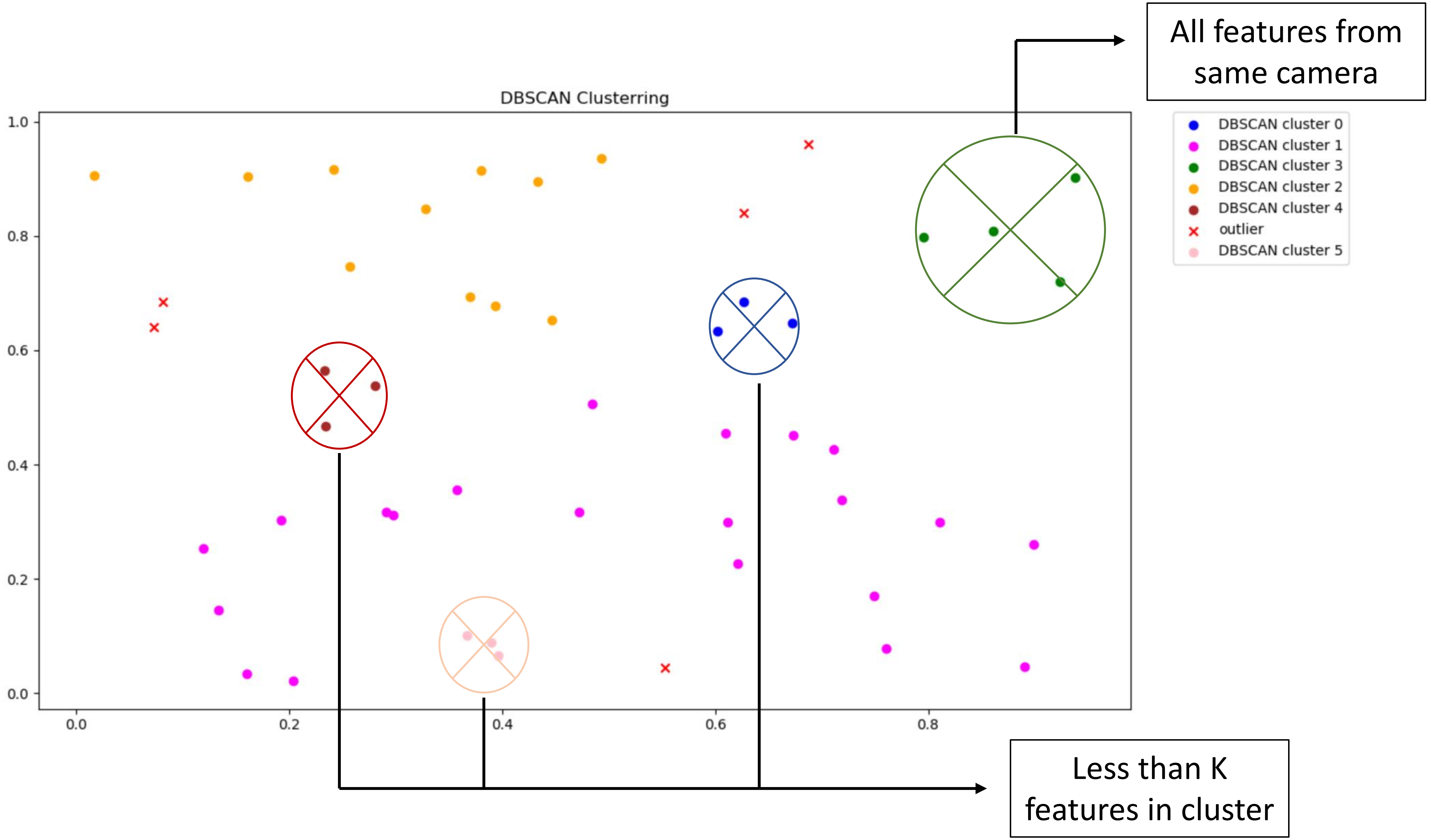


Step 2 – Clustering Technique





Step 3 – Cluster Selection



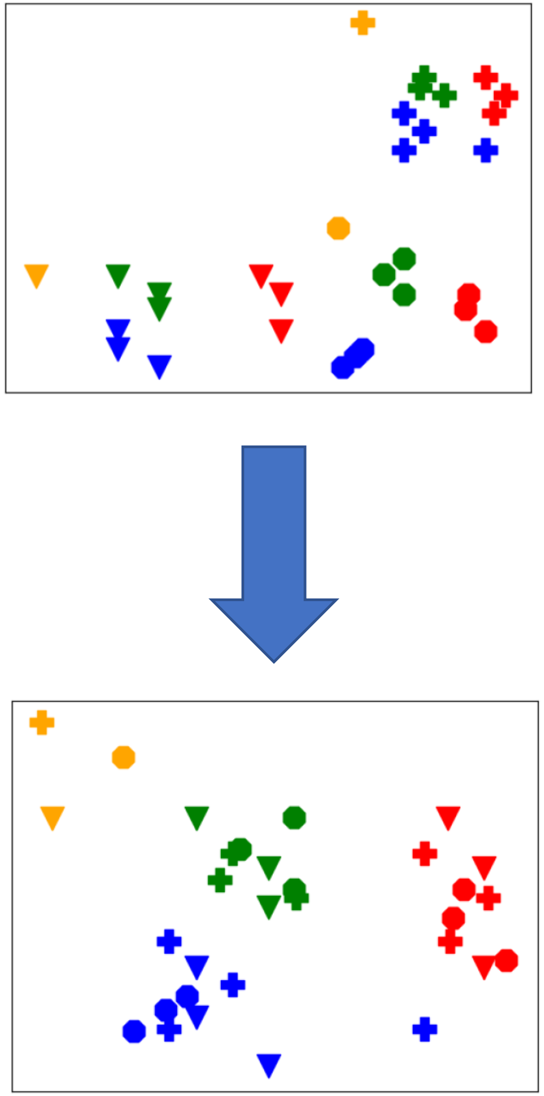
Step 4 – Camera-Guided Feature Normalisation

The high variance present in person Re-ID is mainly caused by different camera views, as each camera has its own characteristics.

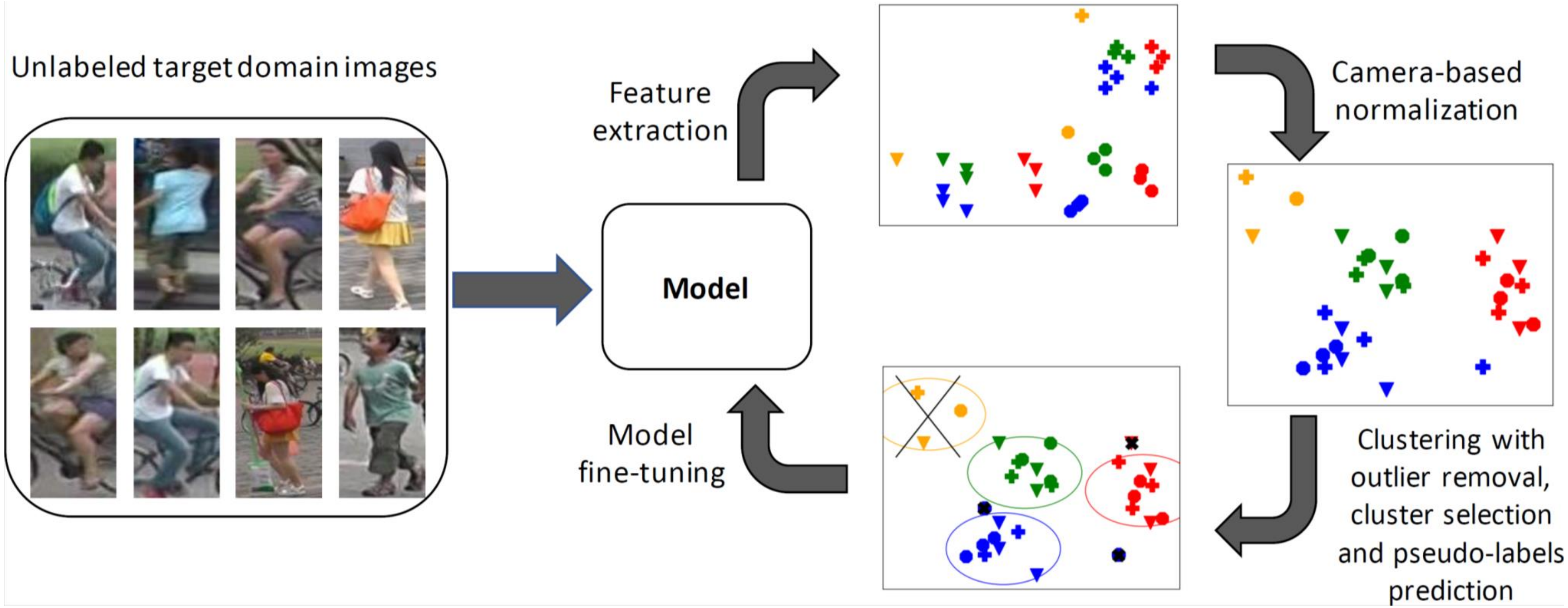
Therefore, the model tends to cluster images by cameras rather than clustering images from the same person in different views.

A camera guided normalisation step is then necessary to reduce this variance and allow the clustering step to create better clusters. The normalisation is done by:

$$\bar{f}_{v_j} = \frac{f_{v_j} - \mu_{v_j}}{\sigma_{v_j}}$$



Step 5 – Unsupervised Domain Adaptation



Baseline Results

Supervised Training	Market1501 -> DukeMTMC				DukeMTMC -> Market1501			
	Rank – 1	**Rank – 5**	**Rank - 10**	**mAP**	**Rank – 1**	**Rank – 5**	**Rank - 10**	**mAP**
Source	44.7	60.7	66.4	27.3	58.9	74.3	80.1	29.0
Target	82.7	92.1	94.6	68.6	92.5	97.6	98.7	81.5
Source and Target	83.9	92.5	94.8	71.1	92.6	97.7	98.6	81.2
Source (**Ours**)	82.7	90.5	93.5	69.3	89.1	95.8	97.2	73.6

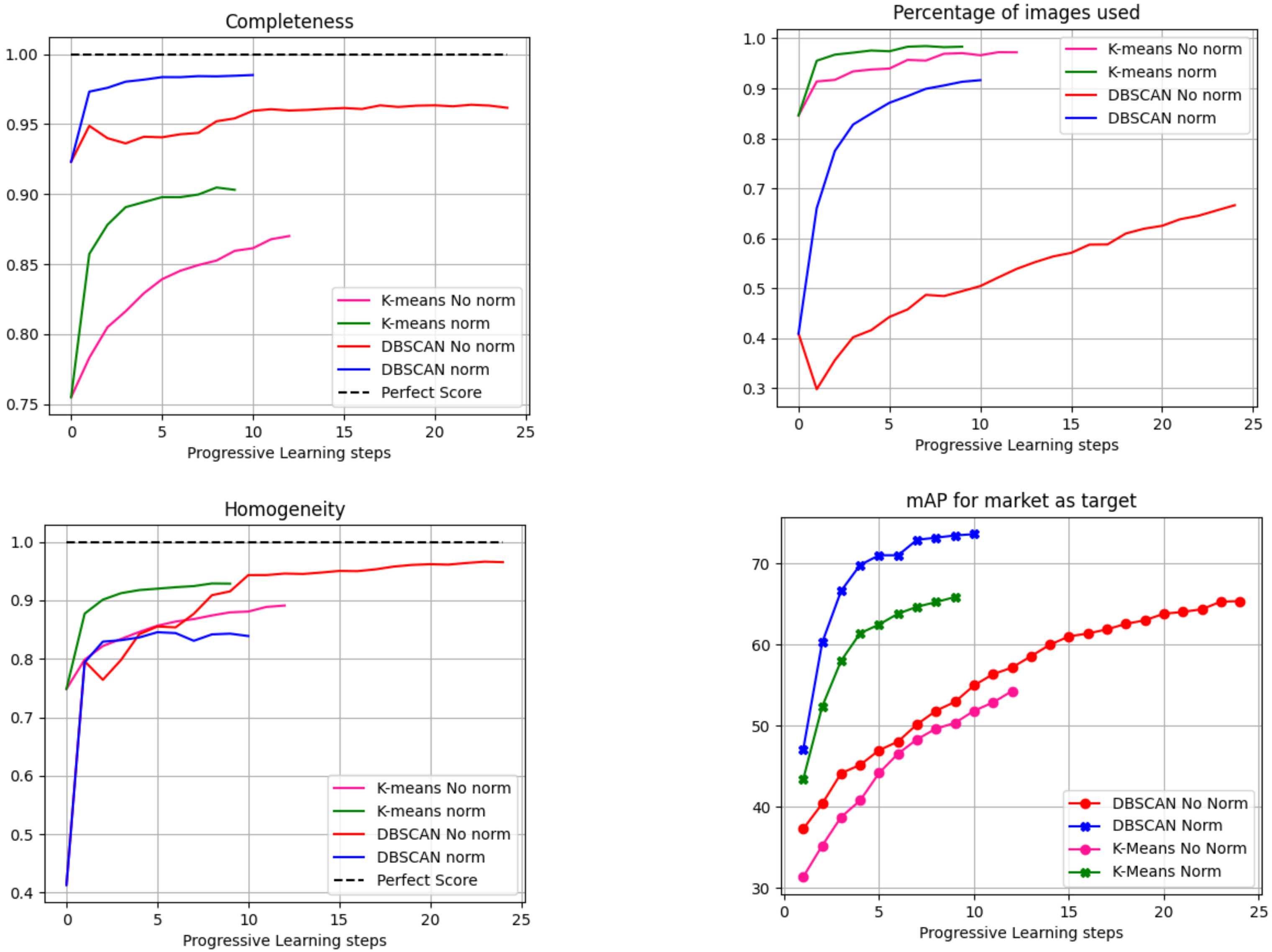
Results

Methods	Market1501 -> DukeMTMC				DukeMTMC -> Market1501			
	Rank – 1	**Rank – 5**	**Rank - 10**	**mAP**	**Rank – 1**	**Rank – 5**	**Rank - 10**	**mAP**
SPGAN	46.9	62.6	68.5	26.4	58.1	76.0	82.7	26.9
UCDA-CCE	55.4	-	-	36.7	64.3	-	-	34.5
ARN	60.2	73.9	79.5	33.4	70.3	80.4	86.3	39.4
MAR	67.1	79.8	-	48.0	67.7	81.9	-	40.0
ECN	63.3	75.8	80.4	40.4	75.1	87.6	91.6	43.0
PDA-Net	63.2	77.0	82.5	45.1	75.2	86.3	90.2	47.6
EANet	67.7	-	-	48.0	78.0	-	-	51.6
CBN + ECN	68.0	80.0	83.9	44.9	81.7	91.9	94.7	52.0
Theory	68.4	80.1	83.5	49.0	75.8	89.5	93.2	53.7
CR-GAN	68.9	80.2	84.7	48.6	77.7	89.7	92.7	54.0
PCB-PAST	72.4	-	-	54.3	78.4	-	-	54.6
AD Cluster	72.6	82.5	85.5	54.1	86.7	94.4	96.5	68.3
SSG	76.0	85.8	89.3	60.3	86.2	94.6	96.5	68.7
DG-Net++	78.9	87.8	90.4	63.8	82.1	90.2	92.7	61.7
MMT	*79.3*	*89.1*	*92.4*	*65.7*	<u>90.9</u>	**96.4**	**97.9**	<u>76.5</u>
Ours	<u>82.7</u>	<u>90.5</u>	**93.5**	<u>69.3</u>	*89.1*	<u>95.8</u>	<u>97.2</u>	*73.6*
Ours + RR	**84.8**	**90.8**	<u>93.2</u>	**81.2**	**92.0**	*95.3*	*96.6*	**88.1**

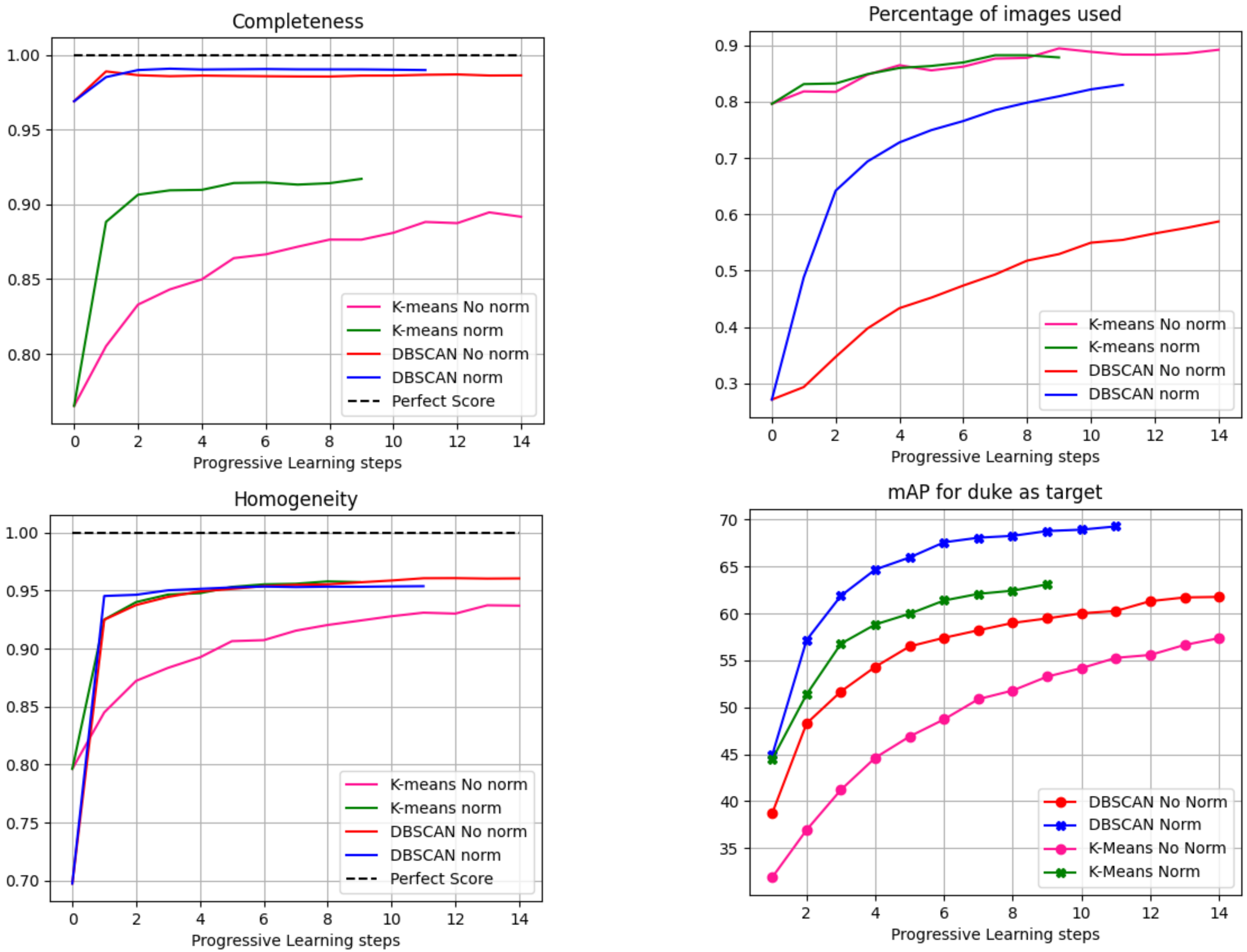
Results – Ablation Studies

Methods	Market1501 -> DukeMTMC		DukeMTMC -> Market1501	
	Rank – 1	**mAP**	**Rank – 1**	**mAP**
Resnet-50	41.4	25.7	54.3	25.5
+ IBN-Net50-a	44.7	27.3	58.9	29.0
+ Domain Adaptation	52.2	37.1	60.1	34.8
+ Progressive Learning	52.2	27.1	61.4	35.5
+ Cluster Selection	77.2	61.8	86.5	66.0
+ Camera Normalisation	82.7	69.3	89.1	73.6

Results – Cluster Methods Comparison – Market1501



Results – Cluster Methods Comparison – DukeMTMC



Final Observations

In this work we proposed two methods for unsupervised domain adaptation in person Re-ID:

- The GAN + pseudo-labels method, which created a baseline for us with the resnet-50 and proved its efficiency with the addiction of the AlignedReID++

- The Multi-Step Pseudo-Labels Refinement that solved some flaws present in the previous method and achieved a state-of-the-art result

Future Work

Although our results are very satisfactory, there is still room for improvement in the UDA Person Re-ID area. We believe that the above ideas are promising:

- Using self-supervised methods (e.g. [15][16][17]) to warm up the initial model direct in the target domain instead of relying in a labeled source domain;
- Using newer neural networks architectures as backbone (e.g., Swin Transformers [18] and ConvNeXt [19]).

[15] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. ICML, 2020.
[16] Chen, X. and He, K.: Exploring Simple Siamese Representation Learning. CVPR, 2021.
[17] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020.
[18] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ArXiv, 2021.
[19] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., and Xie, S.: A ConvNet for the 2020s ArXiv, 2022.

Thank You!