Person Re-Identification in a Video Sequence

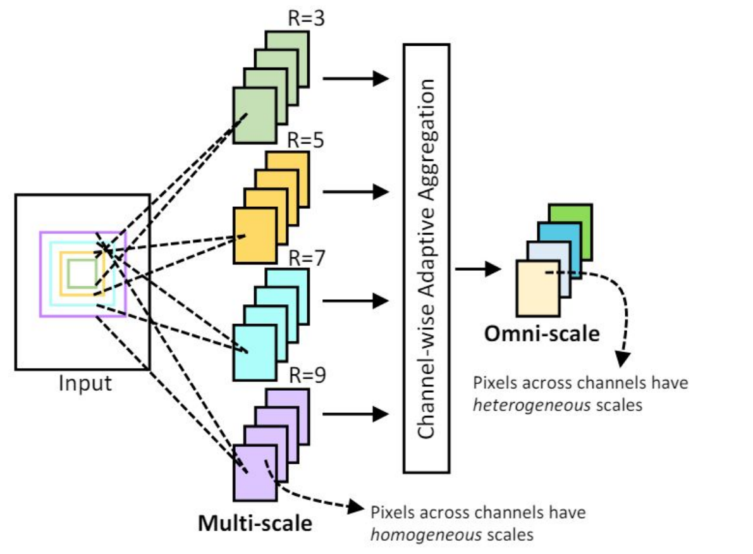
Jiayang Wang[1], Zhiyuan Li[1]

Stanford University

Stanford
School of Engineering

Introduction

Person re-identification is a computer vision task that aims to detect and identify a person of interest with various poses and orientations across different locations. In this paper, we train a CNN-based network, OSNet [1][2], on two popular datasets, and apply the model to build an application that can find target persons in an input video.


R=3
R=5
R=7
R=9
Input
Multi-scale
Channel-wise Adaptive Aggregation
Omni-scale
Pixels across channels have *heterogeneous* scales
Pixels across channels have *homogeneous* scales

A schematic of the proposed building block for OSNet. R: Receptive field size. [1][2]

Sample video that contains walking pedestrians in daylight.

Datasets

In this project, we utilize two datasets: Market-1501 and DukeMTMC-reID.

Market-1501:

- 32,000+ images of 1,501 identities from 6 cameras (128x64 pixels), divided into 12,936 training and 19,732 testing images.e

DukeMTMC-reID:

- 36,000+ images of 1,404 identities from 8 cameras (128x64 pixels), divided into 16,522 training and 17,661 testing images.

DukeMTMC-reID is more challenging due to its greater variability in scenes, lighting, and occlusions, while Market-1501 offers a more uniform environment with slightly less variation.

The images depict pedestrians captured from various camera angles within the Market-1501 dataset]. In each pair of adjacent images, the pedestrians shown belong to the same identity. The images include sample pedestrians from the DukeMTMC-reID dataset, where each set of adjacent images in both rows and every five columns represents the same identity.

Methods & Experiments

Refinement of OSNet with Two-stepped Transfer Learning

OSNet performs well in terms of R1 and mAP but struggles when tested on different datasets. For example, OSNet trained on Market-1501 only achieves an R1 of 30.1% and mAP of 15.6% on DukeMTMC-reID. Combining datasets for training is impractical due to high computational demands.

We use Two-stepped Transfer Learning to address this issue. This method involves:

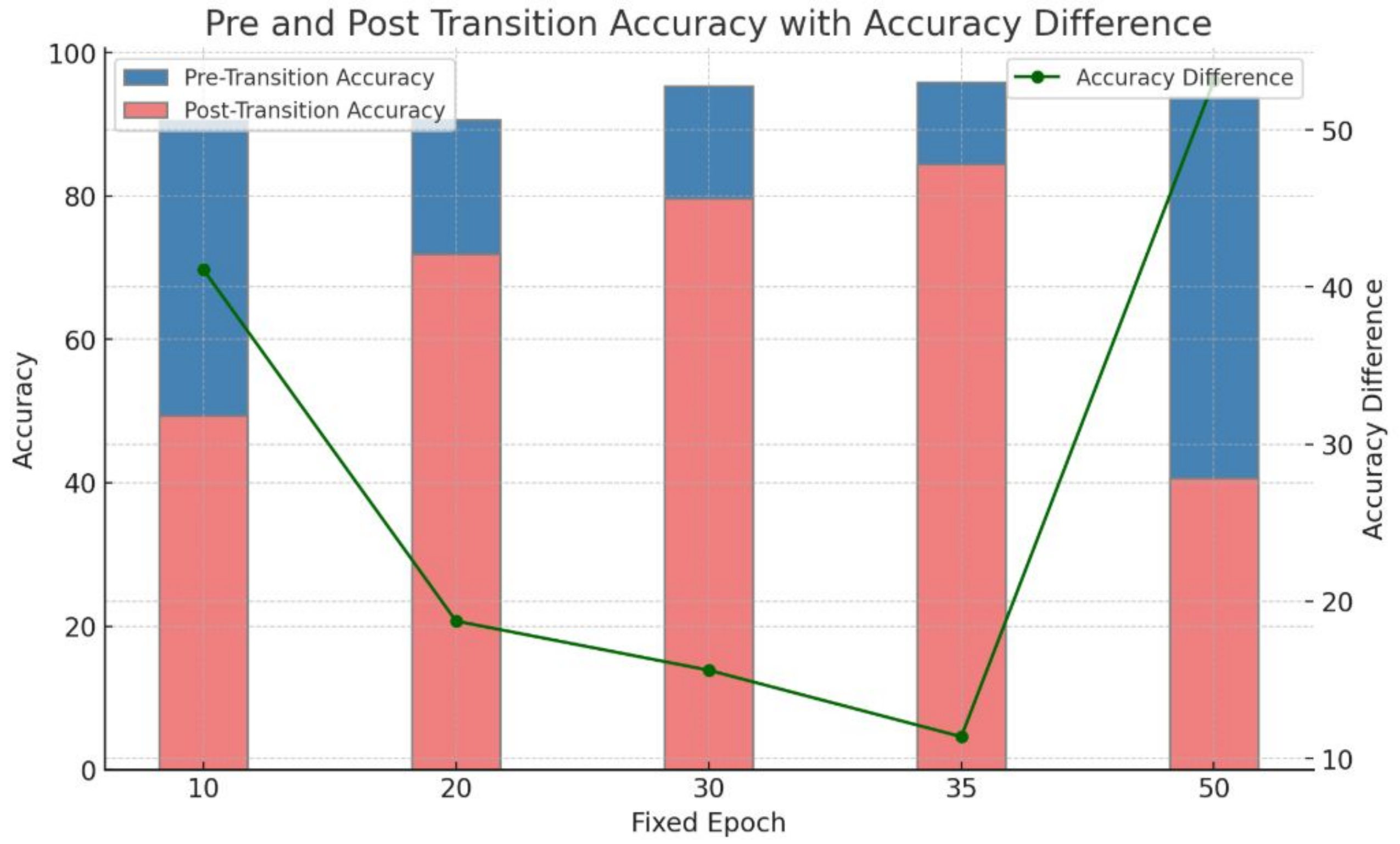
1. Freezing the base layers during initial training for a few epochs.
2. Unfreezing them after pre-training the randomly initialized layers at the network's end.

This approach preevents the pre-trained model from being affected by unfavorable gradients from the new layers, ensuring a smooth transition between datasets. We use 35 fixed epochs for a controlled transition from regional to global training, minimizing performance drops.

Fixed Epoch	Pre-Transition Accuracy	Post-Transition Accuracy	Accuracy Drop
10	90.4883	49.3750	41.1133
20	90.6250	71.8750	18.7500
30	95.3125	79.6875	15.6250
35	95.8464	84.4531	11.3933
50	93.7500	40.6250	53.1250

Impact of Fixed Epochs on Accuracy During Transition Stage. This table presents the pre-transition and post-transition accuracies for different fixed epochs during two stepped transfer learning. The accuracy difference highlights the drop observed at the transient stage from frozen to unfrozen layers.

```
Algorithm 1: Identify person queried in the current frame
  Input: frame, query
  Output: detection
  Input  : Picture frame, Query image
  Output: The same person in the query image that
          appears in the frame
  queryFeature ← OSNet(query);
  queryFeature ← Normalize(queryFeature);
  personBoxes ← YOLO(frame);
  foreach $person$ in $personBoxes$ do
      personFeature ← OSNet(person);
      personFeature ← Normalize(personFeature);
      distance ← queryFeature · personFeature;
      if $distance < threshold$ then
          detection ← person;
          break;
  return $detection$
```



Comparison of Pre-Transition and Post-Transition Accuracies Across Different Fixed Epochs in Fine-Tuning Stage. The bars represent the accuracy before (steel blue) and after (light coral) unfreezing the base layers. The green line indicates the accuracy difference, highlighting the impact of insufficient adaptation when the base layers are fixed for an extended period

References

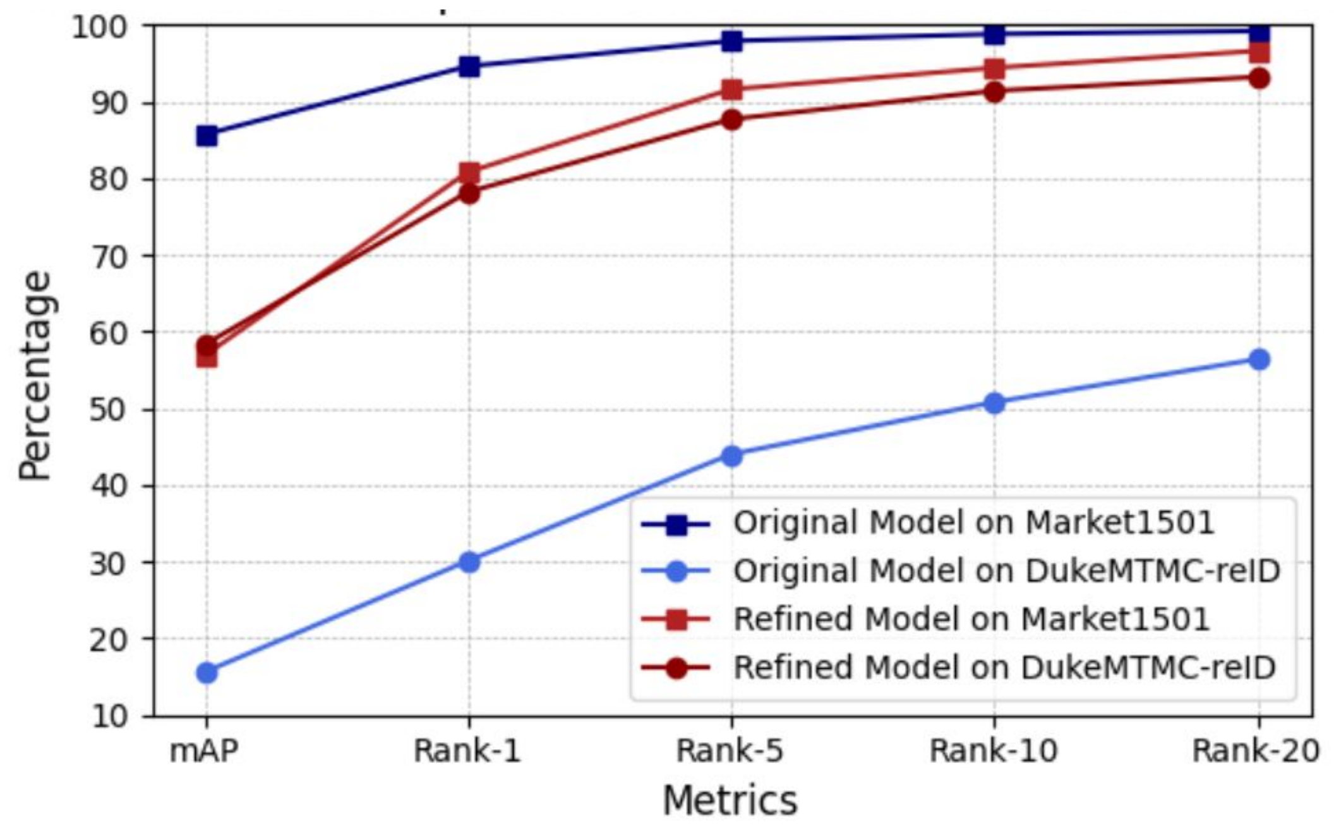
[1] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In ICCV, 2019.

[2] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Learning generalisable omni-scale representations for person reidentification. TPAMI, 2021.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

Results

Performance Comparison between two datasets:



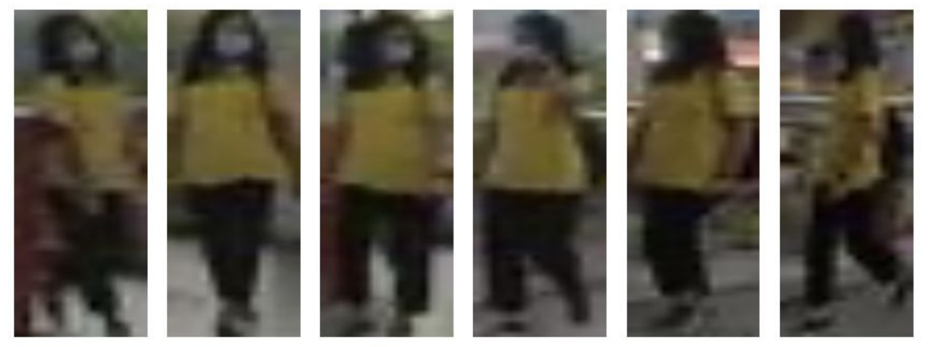
The graph compares the performance of original and refined models on Market1501 and DukeMTMCreID daetasets using metrics mAP, Rank-1, Rank-5, Rank10, and Rank-20. Navy blue (Market1501) and royal blue (DukeMTMC-reID) lines represent the original model, while firebrick (Market1501) and dark red (DukeMTMCreID) lines represent the refined model. The refined model significantly improves on DukeMTMC-reID and achieves balanced performance across both datasets. The x-axis shows the metrics, and the y-axis indicates their percentage values.

Detection Results from the video :

Query A Query B Query C Query D

Queried Pedestrian Targets

Query A Detected In Video



Query B Detected In Video

Query C Detected In Video

Query D Detected In Video

Query	Detections	Total Target Appearances	Rate of Success	Number of False Positives
A	6	16	38%	0
B	99	244	41%	5
C	182	318	57%	0
D	42	49	86%	0

Query	Detections	Total Target Appearances	Rate of Success	Number of False Positives
A	9	16	56%	0
B	201	244	82%	2
C	163	318	51%	0
D	37	49	76%	0

Comparison Between Detection Results By Original Model (Left) and Refined Model (Right)