Person Re-identification

Nguyen Minh Bao Huy
22127155
nmbhuy22@clc.fitus.edu.vn

Ton That Minh Quan
22127349
ttmquan221@clc.fitus.edu.vn

***Abstract*—Person re-identification (re-ID) aims to match images of the same individual across non-overlapping camera views, using pedestrian detection as its foundation. This instance-level recognition task faces significant challenges including variations in illumination, pose, viewpoint, motion blur, resolution, camera configurations, occlusions, and cluttered backgrounds. To address these complexities, re-ID systems rely on highly discriminative features capable of capturing and adaptively combining multi-scale spatial information. We call features of both homogeneous and heterogeneous scales omni-scale features.**

In this paper, we propose OSNet [6], a deep CNN architecture for re-ID that enables omni-scale feature learning. The proposed solution implements a multi-stream residual block architecture, with each convolutional stream dedicated to feature extraction at a particular scale. Central to this design is a novel unified aggregation gate that dynamically merges multi-scale features through learnable, input-dependent channel weights. For computational efficiency and regularization, the block leverages pointwise and depthwise convolutional operations to capture spatial-channel correlations.

I. Introduction

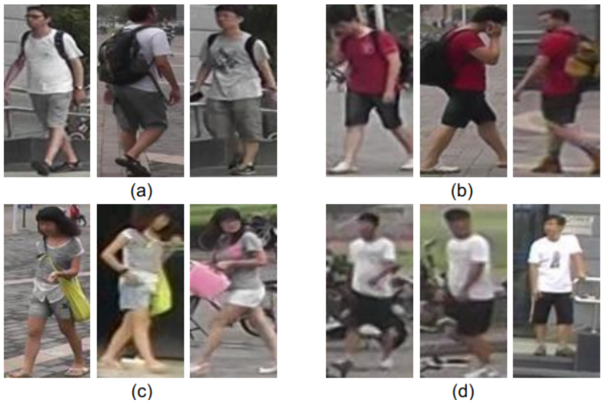

Figure 1. Person re-ID is a hard problem, as exemplified by the four triplets of images above. Each sub-figure shows, from left to right, the query image, a true match and an impostor/false match.

A. Motivation

The goal of person re-identification is to identify and match pedestrians captured by different, non-overlapping cameras using visual characteristics. However, matching the visual features of pedestrians captured from different camera angles is extremely challenging due to various factors that cause differences between images, such as variations in lighting, walking posture, image resolution, and so on. Some examples are shown in Figure 1. As a fundamental task in surveillance systems, ReID enables cross-camera tracking for security applications like criminal investigation and preventive monitoring, while also supporting urban management through crowd flow analysis and anomaly detection. As an instance-level recognition task, re-ID faces two major challenges (Fig. 1):

- **Large intra-class variations**: Differences in camera viewpoints and conditions cause significant appearance changes for the same person. For instance, in Figs. 1(a) and 1(b), the backpack's appearance varies drastically due to perspective shifts (frontal vs. back view), complicating matching.
- **Small inter-class variations**: Pedestrians often wear similar clothing, and surveillance footage's low resolution can make distinct individuals appear nearly identical (see the impostors in Fig. 1).

B. Objective

Current person re-identification approaches typically comprise two key elements: (1) feature extraction techniques that encode distinguishing visual patterns from input images, and (2) similarity metrics that quantify feature correspondence across different views. The field's research directions primarily concentrate on three avenues: enhancing discriminative feature representations, developing more effective distance metrics, or jointly optimizing both components. Feature engineering efforts aim to create robust descriptors that maintain stability against photometric and geometric variations, including changes in illumination, body pose, and camera viewpoint. Meanwhile, metric learning strategies seek to project raw features into an optimized embedding space where positive pairs (same identity) exhibit smaller distances than negative pairs (different identities) according to a learned transformation.

C. Problem statement

A typical system takes two full-body images as input and outputs either a similarity score or a binary classification ("same" or "different" identity).

D. Our approach

To address challenges mentioned in I-A, the core of effective re-ID lies in learning discriminative features. We propose that such features must possess an *omni-scale* property—integrating both variable homogeneous scales (uniform-sized features) and heterogeneous scales (multi-scale combinations).

In this paper, we present **OSNet**, a lightweight CNN architecture for omni-scale feature learning in person re-identification. The key innovations include: (1) Multi-stream blocks with exponentially growing receptive fields and dynamic fusion via an Aggregation Gate (AG); (2) Factorized convolutions yielding a model 10× smaller than ResNet50.

II. Overview of Methods

A. Deep ReID - FPNN

[1] The Deep ReID Filter Pairing Neural Network (FPNN) represents a significant advancement in person re-identification by addressing fundamental limitations of traditional approaches. Unlike handcrafted features that struggle to balance discriminative power and robustness, FPNN automatically learns optimal visual features end-to-end while jointly modeling complex cross-camera transformations. Its novel architecture introduces specialized layers - including filter pairs with maxout grouping for photometric adaptation and patch matching layers for geometric alignment - that explicitly handle multimodal variations in illumination and viewpoint without restrictive unimodal assumptions. Crucially, FPNN co-optimizes all system components (feature learning, transform modeling, and metric computation) in a unified framework, overcoming the suboptimal performance of sequential optimization pipelines. This comprehensive approach yields superior performance, with FPNN outperforming state-of-the-art methods (KISSME, eSDC, and SDALF) in Rank-1 accuracy on the CUHK03 benchmark dataset. The network's inherent robustness to real-world challenges like misalignment, occlusions, and background clutter stems from its learned spatial filtering and holistic feature representation. As the first deep learning solution specifically designed for re-ID's unique cross-view challenges, FPNN moves beyond generic CNNs by incorporating domain-specific mechanisms that surpass both traditional feature engineering and conventional deep vision systems in handling the photometric and geometric complexities of person matching across surveillance cameras.

B. Improved deep learning architecture - Ejaz

[2] The model utilizes tied convolution layers instead of single convolution layers to capture local relationships between two input images, following a structured processing sequence: (1) initial tied convolution and max-pooling layers extract high-level features from each image, (2) a specialized feature comparison layer identifies differences by matching features across neighboring regions, (3) a patch summary layer aggregates difference maps into compact representations, and (4) additional convolution/max-pooling layers capture spatial relationships between differences. This architecture demonstrates superior cross-view matching capability, outperforming FPNN by 34% Rank-1 accuracy on CUHK01 and 37.13% on CUHK03 benchmarks. The system culminates in two fully connected layers with softmax loss for final identity verification.

C. Multi-channel part-based model

[3] The multi-channel parts-based CNN model combines hierarchical feature extraction with an improved triplet loss function to jointly learn global and local body features for robust person re-identification. The architecture employs: (1) a global convolutional layer capturing full-body characteristics, (2) four part-based convolutional layers extracting localized features from key body regions, and (3) five channel-wise fully connected layers integrated through a final fusion layer. **This design achieves superior cross-view representation, outperforming both FPNN and Ejaz methods respectively by 25.8% and 6.2% Rank-1 accuracy on CUHK01 benchmark. The improved triplet loss further refines discriminative feature learning, with the global channel providing contextual cues while part-based layers preserve fine-grained spatial details critical for accurate identity matching.

D. Omni-scale network (OSNet)

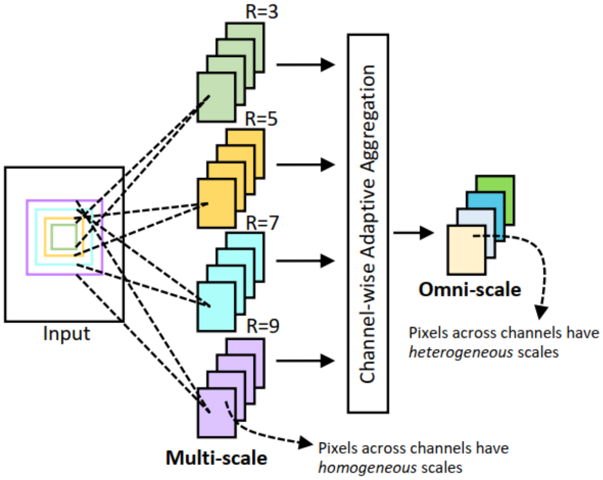
R=3
R=5
R=7
R=9
Input
Channel-wise Adaptive Aggregation
Omni-scale
Pixels across channels have *heterogeneous* scales
Multi-scale
Pixels across channels have *homogeneous* scales

Figure 2. [6]A schematic of the proposed building block for OSNet. R: Receptive field size.

OSNet introduces a novel CNN architecture specifically designed for person re-identification, addressing key limitations of conventional models like ResNet that were developed for category-level recognition. The framework's core innovation lies in its omni-scale feature learning through (1) parallel convolutional streams with exponentially growing receptive fields to capture diverse spatial scales (Fig. 2), and (2) a dynamic aggregation gate (AG) that adaptively fuses features via input-dependent channel weights, enabling both homogeneous (single-scale) and heterogeneous (multi-scale) representations. Unlike prior multi-scale approaches [38,2], OSNet uniquely combines these capabilities while maintaining computational efficiency through factorized (depthwise/pointwise) convolutions, resulting in a model 10× smaller than ResNet50. This lightweight design not only prevents overfitting on moderate-sized re-ID datasets but also enables practical deployment for large-scale surveillance by reducing on-device computation. OSNet achieves state-of-the-art performance across six benchmarks, demonstrating superior accuracy while being significantly more compact than existing solutions.

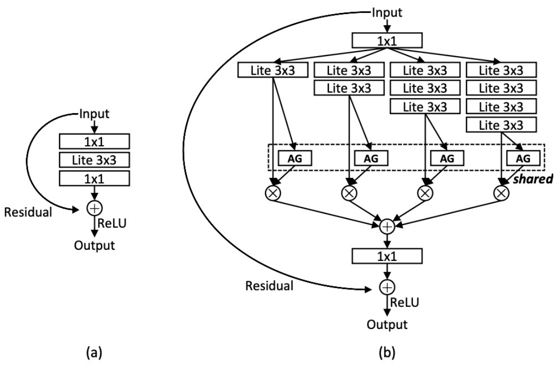

Figure 3. (a)Baseline bottleneck. (b) Proposed bottleneck. AG: Aggregation Gate. The first/last 1x1 layers are used to reduce/restore feature dimension

III. Chosen Method - Omni-Scale Network

The Omni-Scale Network (OSNet) (Fig. 2) is designed to extract omni-scale features, which encompass both homogeneous and heterogeneous scale representations:

- Homogeneous-scale features: Features with identical receptive field sizes.
- Heterogeneous-scale features: Features combining multiple spatial scales, enabling robust handling of viewpoint changes and occlusions through integrated local-global information.

Most CNN architectures, such as ResNet, are designed for category-level recognition rather than instance-level recognition. The Omni-Scale Network is specifically designed to learn Omni-Scale Features through an enhanced residual block derived from the residual bottleneck structure (Fig. 3):

- Multiple convolutional streams: Each branch processes information at different spatial scales.
  - With each branch, we stack t Lite 3x3 convolutions($t > 1$) with receptive field size $(2t+1) \times (2t+1)$. OSNet chooses $t = 4$, meaning that the receptive fields range from 3×3 to 9×9.
  - Shortcut connection preserves features across deeper layers, mitigating information loss.
- Unified Aggregation Gate - UAG: A dynamic feature fusion mechanism that adaptively weights multi-scale branches:
  - UAG uses a small network consist of Global Average Pooling for extracting feature and MLP with ReLU and Sigmoid to create weight for each feature channel.
  - Input-adaptive weighting improves instance-level discrimination, critical for Re-ID tasks.

Omni-Scale Network is a lightweight network:

- In a camera system, it is impossible to transmit video to a central server. The solution is to run a lightweight network on the camera to extract features and transmit back only the necessary information.
- The data for Re-ID is usually moderate in size because it is difficult to collect images of a person from multiple cameras. And the lightweight network to help avoid overfitting.

The model replaces standard convolution with Depthwise Separable Convolution (including Depthwise convolution and Pointwise Convolution) to reduce model size.

- Depthwise convolution processes the space on each individual channel. This class has h.w.c.k.k number of calculations and c.k.k number of parameters.
- Pointwise convolution (1x1) combines multi-channel information. This layer has h.w.c.c' as the number of operations and c.c' as the number of parameters.

So Depthwise Seperable Convolution reduces the number of calculations from h.w.k.k.c.c' to h.w.c.(k.k + c'), reducing the number of parameters from k.k.c.c' to c.(k.k+c').

A. Metrics

In the Person Re-identification (Re-ID) problem, there are two popular evaluation metrics: Cumulative matching characteristics (CMC) Rank-1 accuracy and mAP.

Rank-1 accuracy is the proportion of queries in which the correct image appears at the top of the result list. This metric is suitable when each person has only one image in the gallery. This metric does not reflect the order of the results, but only indicates whether the correct image appears in the top-1 or not.

Mean Average Precision (mAP) evaluates the overall accuracy by considering all the results in the query list. mAP better reflects the system's accuracy on the result list.

B. Loss

In the Person Re-identification (Re-ID) problem, one of the popular loss functions is Cross Entropy Loss. However, when using one-hot labels, the model tends to over-trust the correct label, leading to overfitting. This is a big problem when there are only a limited number of images in each person data. Label Smoothing is a technique that helps the model to be less dependent on the correct label, by blurring the one-hot label instead of setting the correct class to 1.

C. Implementation Details

We implement our model using the Torchreid framework, a PyTorch-based library for deep learning-based person reidentification. All experiments are conducted on the Market-1501 [5] dataset, a widely used benchmark containing 32,668 pedestrian images from 1,501 identities, captured across six cameras.

We adopt OSNet as our backbone network, specifically the lightweight variant osnet_x0_25 to balance accuracy and computational efficiency. The last fully connected (FC) layer outputs 1501 logits. Two training strategies are considered:

- Training from Scratch: The model is initialized with random weights and trained from scratch.
- Pretrained Training: The model is initialized with weights pretrained on ImageNet and fine-tuned on Market-1501.

Model	mAP (%)	Rank-1 (%)
OSNet	18.6	35.5
OSNet+pretrained	44,2	70.1

TABLE I
RESULT ON MARKET-1501 DATASET

Each image is resized to 256×128 pixels. We apply standard augmentation techniques, including:

- Random horizontal flipping with a probability of 0.5,
- Random cropping with zero-padding of 10 pixels,
- Normalization using the ImageNet mean and standard deviation.

The dataset is divided into mini-batches with a batch size of 16 for both training and evaluation.

Our training objective is based on Softmax Cross-Entropy Loss with Label Smoothing. The loss function is formulated as:

We optimize the model using the AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 5×10^{-4}. To further improve convergence, we apply a ReduceLROnPlateau scheduler, which reduces the learning rate by a factor of 0.1 if the validation loss does not improve for 5 consecutive epochs.

We train the model for 10 epochs using the ImageSoftmaxEngine module from Torchreid, with intermediate evaluation at the end of each epoch. All experiments are conducted on Google Colab (using T4 GPU).

D. Result

From Table I, we have the following observations. Our model achieves 18.6% mAP (trained from scratch), 44.2% mAP (fine-tuned from ImageNet), significant gap with Paper's Benchmarks 81% and 84.9%. Rank-1 gives the same result. There are the reasons:

- Computational Resources: Weaker GPU → Reduced Batch Size/Number of Epochs → Poor Convergence and Insufficient Training Data Exposure
- Data Augmentation: Not applying all available augmentation techniques leads to insufficient data diversity.

The model fine-tuned from ImageNet achieves significantly better results. This is because the model has already learned feature extraction from 1,000 classes in the ImageNet dataset, making the learning process much easier and faster.

IV. DATASET MARKET-1501

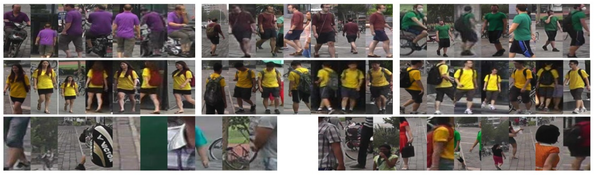

Figure 4. Sample images of the Market-1501 dataset. All images are normalized to 128×64 (Top:) Sample images of three identities with distinctive appearance. (Middle:) We show three cases where three individuals have very similar appearance. (Bottom:) Some samples of the distractor images (left) as well as the junk images (right) are provided.

Market-1501 (Fig. 4) is a large-scale dataset for person re-identification (Re-ID) to facilitate research in pedestrian retrieval. The dataset was collected in front of a supermarket on the Tsinghua University campus using six surveillance cameras, including five high-resolution cameras and one low-resolution camera. The dataset contains 32,668 bounding box images of 1,501 identities, with each identity appearing in at least two different cameras. The images were obtained using the Deformable Part Model (DPM) detector, which introduces misalignment and background clutter, making the dataset more challenging for re-identification tasks.

Market-1501 is divided into a training set and a testing set following a standard split:

- Training Set: 12,936 images of 751 identities.
- Testing Set: 19,732 images of 750 identities, further divided into:
  - Query Set: 3,368 images, each corresponding to a unique person.
  - Gallery Set: 15,913 images, containing multiple instances of each identity, along with junk images (background clutter and detection errors).

REFERENCES

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep Filter Pairing Neural Network for Person Reidentification,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[2] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person reidentification,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3908–3916.

[3] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person Re-identification by Multi-Channel Parts Based CNN with Improved Triplet Loss Function,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1335–1344.

[4] Muna O. AlMasawa, Lamiaa A. Elrefaei (Senior Member, IEEE), Kawthar Moria, ” A Survey on Deep Learning Based Person Re-Identification Systems”, in 2019 IEEE Access, 2019, pp. 175228-175247.

[5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, ”Scalable person re-identification: A benchmark”, in 2015 IEEE International Conference on Computer Vision, 2015, pp. 1116–1124

[6] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, ”Omni-Scale Feature Learning for Person Re-Identification”, in 2019 IEEE International Conference on Computer Vision, 2019, pp. 3702–3711.